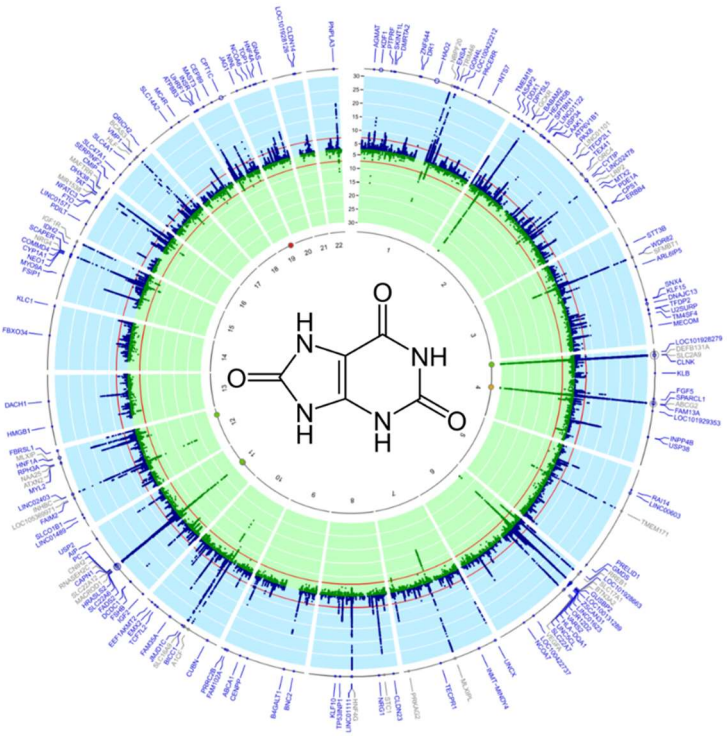
THE UNIVERSITY
of EDINBURGH

The Genetics of Serum Urate Regulation in Human Health and Disease

---

Jonathan Charles Leonard Marten

s1563693



PhD

University of Edinburgh

2019

I. Abstract

Uric acid, the end product of purine catabolism in humans, is a biologically active molecule that plays a role in oxidative stress, inflammation, and the regulation of blood pressure. Excessively high serum urate levels (hyperuricaemia) are associated with a wide range of diseases. With the exception of gout, where monosodium urate crystals are known to trigger a painful inflammatory response, both the causality and the underlying mechanisms linking hyperuricaemia and disease are unclear.

To better understand the link between uric acid and cardiometabolic disease, I have investigated the correlation between serum urate and 266 Olink protein biomarkers associated with cardiovascular disease and inflammation and 191 lipid species. Using partial correlation and lasso regression, I have identified and replicated 11 protein biomarkers whose serum levels covary with urate independently of the other biomarkers. The associated proteins are involved in diverse processes including phosphate metabolism and bone development, glucose metabolism, adipocyte function and blood pressure regulation. I have additionally identified 15 lipids, some of which have a potential link with cognitive function.

To approach the question of uric acid from a regulation perspective, I have run genome-wide association scans, first in our own cohorts, with a sample exceeding 10,000 individuals imputed to the Haplotype Reference Consortium reference panel, the first GWAS of serum urate levels to be run on this panel. Then, as part of the CKDGen consortium, I co-lead a transethnic meta-analysis of over 450,000 individuals, the largest GWAS of serum urate to date. Our work identified 183 urate-associated loci, of which 147 were novel. These loci can be used to create a genetic risk score for urate that has considerable predictive potential for gout, assessed in the UK Biobank.

II. Lay Summary

DNA is composed of two kinds of molecules, purines (As and Gs) and pyrimidines (Cs and Ts). When DNA is broken down into its component parts, surplus purines are disposed of in a complex series of carefully controlled biochemical reactions, the end result of which is a small molecule called uric acid.

Uric acid is much more than just a waste product. It can soak up dangerous free radicals, highly reactive rogue molecules occurring naturally in our bodies which can cause damage if left unchecked. It is also involved in regulating blood pressure and promoting response to damage and infection.

However, the amount you have in your blood needs to be very precisely controlled. Most mammals turn uric acid into another chemical which is easily passed out in urine. Humans have lost the ability to do this, so we have a lot more uric acid in our blood, and if levels get too high, crystals of uric acid can grow. These crystals can cause swelling and intense pain – the disease we know as gout. Gout is becoming increasingly common in developed countries, because we also get too many purines from our diet, particularly red meat, oily fish and alcohol. High uric acid levels are also linked to high blood pressure, heart disease, diabetes, liver disease, kidney disease and obesity. With the exception of gout, we don't yet know whether uric acid is a cause or a consequence of these conditions.

My project aims to help understand how uric acid can lead to disease, and how your genes can affect the amount of uric acid in your blood.

To investigate the first point, I have been looking at several hundred molecules that have functions related to heart disease. I have identified eleven which seem to be closely linked to uric acid levels. With further study, these may help to understand mechanisms connecting uric acid to heart disease.

To address the second question, I have joined an international collaboration looking at small changes in DNA between hundreds of thousands of people. I have helped to identify 183 changes which seem to cause differences in uric acid level. By looking at these changes, we can try to understand what their function is, which may help to understand how uric acid levels are controlled. If we can do this, we can predict people's disease risk from their DNA. It might also be possible to design drugs based on these changes which could be used to treat gout and other diseases.

III. Declaration

I declare that I composed this thesis, and the work herein is my own. Where I have been part of a research consortium, my contribution has been clearly indicated.

This work has not been submitted for any other degree or professional qualification.

Jonathan Marten

25th April 2019

IV. Acknowledgements

First and foremost, I want to thank to my supervisor Veronique Vitart, whose patient guidance was indispensable at every step of my PhD. I would also like to thank my second supervisor, Caroline Hayward, for her help and advice.

I would like to thank the participants and organisers of the CROATIA-Vis, CROATIA-Korcula, CROATIA-Split, ORCADES, VIKING and Generation Scotland: Scottish Family Health Study cohorts for the use of their data, and the INTERVAL, EGCUT, PIVUS and Lifelines DEEP cohorts for replication. Analyses in this thesis were performed variously under UK Biobank projects 8304, 12611, 19655 and 20272. I would also like to thank the CKDGen consortium, in particular Anna Köttgen and Adrienne Tin, who provided guidance and gave me many opportunities to carve out my own role in the project. From the University of Edinburgh, I would like to thank Athina Spiliopolou for help with GENOSCORES and Andy Bretherick, Anne Richmond and Thibaud Boutin for the Olink meta-analyses.

I am indebted to the QTL group for keeping me fuelled on cake and tolerating my grumpiness. Thanks also to former group members Réka Nagy and Jenny Huffman for moral support, guidance and answers to stupid questions.

I want to thank my friend Karen Vincent for being a sympathetic ear, an occasional welcome distraction and the unofficial project manager for my thesis, without whom I would still be trying to write the first chapter.

My partner, Lisa Ackerman, never once wavered in her support and encouragement. She always knew how to help me through and made even the hardest days worth it. I love you so much.

I also want to thank my family: my brother Graham, who was always good at science and made me want to be too, and my sister Rachel, who kept me hooting with laughter at inconvenient times with jokes that no one else understood. I especially want to thank my parents. Mum and Dad, you made me who I am today and gave me all the love and support a son could want. I couldn't have done this without you. Thanks also to Pippin, who kept me company while I was writing over Christmas, even if it was mostly to see if I'd feed him a second lunch. I love you all very much.

Finally, to my grandfather: Grandpa, you were so proud when I told you I was going to start a PhD. I know that you'd have been even more proud to see me finish.

V. Table of Contents

I. Abstract 3
II. Lay Summary 5
III. Declaration 7
IV. Acknowledgements 9
V. Table of Contents 11
VI. List of Tables 15
VII. List of Figures 17
VIII. List of Abbreviations 19
IX. Note on genomic positions 21
Chapter 1 Introduction 23
1.1 Uric acid 23
1.1.1 What is uric acid? 23
1.1.2 Urate production 25
1.1.3 Properties of uric acid 28
1.1.4 Uric acid transport 31
1.1.5 Effect of sex hormones 33
1.1.6 Genome-wide association studies of serum urate 34
1.2 Clinical and epidemiological relevance 37
1.2.1 Gout 38
1.2.2 Cardiometabolic disease 39
1.2.3 Renal disease 42
1.2.4 Neurodegenerative disease 43
1.3 Project Aims 43
Chapter 2 Phenotype Correlations 45
2.1 Background 45
2.1.1 Cohorts 48
2.1.2 Phenotypes 50
2.2 Methods 51
2.2.1 Data processing 51
2.2.2 Correlations 56
2.2.3 Partial correlations 56
2.2.4 Lasso regression 57
2.2.5 Replication 58
2.2.6 Serum urate GWAS lookups 59
2.2.7 GENOSCORES 61
2.2.8 Genetic correlation with LD-score regression 62

2.3 Olink Results 63
2.3.1 Cohort summary 63
2.3.2 Correlations 64
2.3.3 Partial correlations 68
2.3.4 Lasso regression 75
2.3.5 Serum urate GWAS lookups 77
2.3.6 GENOSCORES 85
2.3.7 Genetic correlation with LD-score regression 87
2.4 Lipidomics results 89
2.4.1 Analysis sample sizes 89
2.4.2 Correlations 89
2.4.3 Partial Correlations 92
2.4.4 Lasso regression analysis 93
2.5 Combined analysis 94
2.6 Conclusions 95
2.6.1 FGF-23 97
2.6.2 FGF-21 99
2.6.3 IGFBP-2 100
2.6.4 Ep-CAM 100
2.6.5 Renin 101
2.6.6 LDL-receptor 102
2.6.7 FABP4 102
2.6.8 IL-1RA 103
2.6.9 CCL3 104
2.6.10 PLC 104
2.6.11 PON3 104
2.6.12 Lipid associations 105
2.6.13 Summary 106
Chapter 3 GWAS of serum urate 109
3.1 Background 109
3.1.1 Genome-wide association studies 109
3.1.2 GWAS of serum urate 111
3.2 Methods 113
3.2.1 Cohorts 113
3.2.2 Genotype imputation 115
3.2.3 Regression analysis 115
3.2.4 Conditional analysis 116
3.3 Results 116

3.3.1 Cohort summary information....116
3.3.2 GS Electronic Health Records....117
3.3.3 Meta-analysis....119
3.3.4 Conditional analysis....127
3.4 Conclusions....129
3.4.1 GS Tayside EHR linkage....129
3.4.2 Serum urate meta-analysis....129
Chapter 4 CKDGen Consortium meta-analysis of serum urate....133
4.1 Background....133
4.1.1 GWAS consortia....133
4.1.2 The CKDGen consortium....134
4.1.3 A note on contributions....134
4.2 Methods....137
4.2.1 Study level analysis & QC....137
4.2.2 Meta-analysis....138
4.2.3 Trans-ethnic meta-regression in MR-MEGA....140
4.2.4 Conditional Analysis in GCTA....143
4.2.5 Genetic risk score and gout in UK Biobank....145
4.2.6 DEPICT pathway analysis....146
4.2.7 FUMA gene function annotation....148
4.2.8 Genetic correlations....149
4.3 Results....150
4.3.1 Summary of participants....150
4.3.2 Transethnic meta-analysis....150
4.3.3 Ancestry-specific analyses....153
4.3.4 Trans-ethnic meta-regression in MR-MEGA....153
4.3.5 Sex-stratified analysis....157
4.3.6 Lookup of Chapter 3 SNPs....161
4.3.7 Genetic risk score and gout in UK Biobank....162
4.3.8 DEPICT Pathway Analysis....165
4.3.9 FUMA gene function annotation....169
4.3.10 Genetic correlations....171
4.4 Conclusions....176
4.4.1 Meta-analysis....176
4.4.2 Trans-ethnic meta-regression in MR-MEGA....178
4.4.3 Sex-stratified analysis....181
4.4.4 Serum urate GRS and Gout....182

4.4.5 Genetic correlations between serum urate levels and published GWAS phenotypes 183
4.4.6 Limitations of annotation 184
4.4.7 The 'omnigenic' hypothesis 185
4.4.8 The future of GWAS 186
Chapter 5 Gout risk in the UK Biobank 189
5.1 Background 189
5.1.1 The UK Biobank cohort 189
5.1.2 Gout case/hypernormal control analysis 190
5.2 Methods 191
5.2.1 Generation Scotland 191
5.2.2 Phenotypes 192
5.2.3 Linear regression 192
5.2.4 Gout risk score in UK Biobank 192
5.2.5 Comparison to serum urate risk score 193
5.3 Results 193
5.3.1 Generation Scotland phenotype summaries 193
5.3.2 Gout risk model 194
5.3.3 Correlation with serum urate in Generation Scotland 194
5.3.4 UK Biobank phenotypes 195
5.3.5 UK Biobank risk scores 196
5.3.6 Relationship to serum urate GRS 198
5.3.7 Defining hypernormals 198
5.4 Conclusions 199
Chapter 6 Conclusions and summary 201
X. References 207
XI. Supplementary Tables 217

VI. List of Tables

Table 1 - Olink proteins with more than 5% of samples set to LLOD before filtering. 52
Table 2 – Sample sizes and phenotype summary statistics for serum urate-Olink correlation analyses. 63
Table 3 – Non-Olink phenotype summary statistics. 64
Table 4 – Phenotypes with significantly different correlations with SUA between sexes (n_{female} = 1083, n_{male} = 809) 65
Table 5 – Partial correlations in the discovery sample. 68
Table 6 – pQTLs identified in each dataset 78
Table 7 – Lookup in CKDGen meta-analysis of serum urate (see Chapter 4) of pQTL Index SNPs from Vis & ORCADES, OLINK-IMPROVE and INTERVAL. 82
Table 8 – Vis & ORCADES Index SNPs from suggestively significant loci (P < 1 x 10^{-5}), with a significant association with serum urate (P < 0.05/467). ... 83
Table 9 – Colocalisation results for rs780094 and rs799167. 83
Table 10 – All GENOSCORES regions with a serum urate-Olink correlation > 0.1. 86
Table 11 – Serum urate-Olink genetic correlations from LDSC. 88
Table 12 – Sample sizes for serum urate-lipidomic correlation analyses. 89
Table 13 – Serum urate-lipidomic correlations with a significant difference in effect between sexes 90
Table 14 – Partial correlation coefficients for BMI and eGFR in serum urate-lipid models. 92
Table 15 – Per-cohort summary statistics 117
Table 16 – Genome-wide significant index SNPs from the GS Tayside serum urate GWAS. 119
Table 17 – Index SNPs from the sex-combined and sex-separate meta-analyses. 123
Table 18 – Lookup of GS Tayside novel loci in meta-analysis results. 125
Table 19 – Sample sizes for serum urate and gout meta-analyses 150
Table 20 – Ancestry-specific loci not identified in the transethnic meta-analysis. .. 153
Table 21 – METAL index SNPs with significant ancestry-associated heterogeneity ($P_{anc-het}$ < 0.05 / 183) 155
Table 22 – MR-MEGA loci which did not contain a METAL index SNP 155

Table 23 – Lookups of MR-MEGA unique index SNPs in ethnic-specific meta-analyses. 156
Table 24 – Female-specific loci for serum urate. 158
Table 25 – SNPs with suggestive sex-effect differences ($P_{diff} < 10^{-5}$) 160
Table 26 – Lookup of Chapter 3 novel SNPs in CKDGen meta-analyses 160
Table 27 – Study-specific GWAS summary statistics for rs573624409. 162
Table 28 – Summary of serum urate GRS bin demographics and logistic regression model. 163
Table 29 – DEPICT Exemplar Gene Sets. 166
Table 30 – Traits with a nominally significant difference in genetic correlation between sexes. 174
Table 31 – Summary of quantitative phenotypes tested for gout risk model in Generation Scotland 193
Table 32 – Coefficients included in the gout risk score model 194
Table 33 – Summary of quantitative phenotypes used to generate gout risk score in UK Biobank. 196

Supplementary Tables

Supplementary Table 1 – List of all Olink proteins and non-Olink phenotypes included in correlation and lasso regression analyses. 217
Supplementary Table 2 – Serum urate-Olink lasso regression mean coefficients. 220
Supplementary Table 3 - Serum urate-lipidomic lasso regression mean coefficients. 221
Supplementary Table 4 – 183 index SNPs identified in the CKDGen trans-ethnic meta-analysis 223
Supplementary Table 5 – MR-MEGA SNPs not identified as METAL index SNPs with the filter on number of cohorts > 37 relaxed. 227
Supplementary Table 6 – Spearman correlations between DEPICT exemplar gene sets ($r > 0.2$) 228
Supplementary Table 7 – All sex-separate genetic correlations significant in males or females. 231

VII. List of Figures

Figure 1 - Chemical structure of uric acid. 23
Figure 2 - Uric acid pathways. 26
Figure 3 - Antioxidant and prooxidant effects of uric acid. 30
Figure 4 - Uric acid transporters in renal proximal epithelial cells. 33
Figure 5 – Mechanisms of metabolic syndrome. 41
Figure 6 - An example of the technical bias visible in the lipidomics measurements. 53
Figure 7 – Comparison of mean standard errors across batch sizes. 55
Figure 8 - Heatmap of correlation coefficients between all phenotypes 66
Figure 9 - Heatmap of Spearman's rank correlations between serum urate and Olink phenotypes. 67
Figure 10 - Partial correlation networks 70
Figure 11 – Partial correlation sensitivity analysis results for whole discovery. 72
Figure 12 – Partial correlation replication. 74
Figure 13 - Mean lasso regression coefficients. 76
Figure 14 - Mean lasso coefficients for phenotypes included in >95% of models, per cohort. 77
Figure 15 – Distribution of index SNPs from Olink meta-analyses of Vis & ORCADES. 81
Figure 16 – GENOSCORES score-score correlation plot for serum urate and Olink loci. 87
Figure 17 - Heatmap of serum urate-lipidomic correlations. 91
Figure 18 - Partial correlation coefficients for BMI and eGFR in serum urate-lipid models. 93
Figure 19 – Mean lasso regression coefficients for serum urate-lipidomics analysis. 94
Figure 20 - Mean lasso coefficients for serum urate regressed on all phenotypes. 95
Figure 21 - Possible mechanisms driving an association between serum urate (U) and omic biomarker (O) levels where a shared genetic variant (G) can be identified. 97
Figure 22 - Distribution of serum urate measurements in GS EHR records compared to all cohorts with directly-measured serum urate. 118
Figure 23 - Miami plot of GS Tayside uric acid results 119

Figure 24 - Manhattan plots of uric acid meta-GWAS. 122
Figure 25 - Forest plots of novel index SNPs with I2 < 60%. 124
Figure 26 - Forest plot of novel index SNPs from GS Tayside GWAS. 125
Figure 27 – Per-cohort Chromosome 11 Manhattan plots. 126
Figure 28 – Manhattan plots for chromosome 11 of conditional analysis on rs542534688. 128
Figure 29 - Overview of the CKDGen serum urate analysis workflow. 136
Figure 30 - Principal component plots from MR-MEGA 143
Figure 31 - Circos plot summarising the sex-combined trans-ethnic meta-analysis results. 152
Figure 32 - Manhattan plot of the p-values for effect size differences between females and males. 159
Figure 33 - Forest plot for rs573624409 in CKDGen cohorts. 162
Figure 34 - Serum urate genetic risk score and gout prediction 164
Figure 35 - Correlation network of exemplar gene sets from DEPICT pathway analysis. 168
Figure 36 – FUMA-generated heatmap of GTEx gene expression for the serum urate loci. 170
Figure 37 – FUMA-generated plot of differentially-expressed genes by tissue type. 171
Figure 38 – Genetic correlations between serum urate EA and LD-Hub traits 172
Figure 39 - Comparison of genetic correlations with serum urate between males and females. 175
Figure 40 - Scatterplot of gout phenotypic risk score against serum urate level in Generation Scotland..... 195
Figure 41 - Gout risk score distribution in UK Biobank. 197
Figure 42 - ROC curve for gout status prediction in UK Biobank from gout risk score. 197
Figure 43 - Scatterplot of gout risk score against serum urate genetic risk score in the UK Biobank. 198

VIII. List of Abbreviations

1000G – The 1,000 Genomes Project

AA – African American ancestry

AIC – Akaike information criterion

ANOVA – Analysis of variance

AMP – Adenosine monophosphate

ATP – Adenosine triphosphate

AUC – Area under the curve

BIC – Bayesian information criterion

BMI – Body mass index

BMD – Bone mineral density

CHD – Coronary heart disease

CKD – Chronic kidney disease

CKDGen – Chronic Kidney Disease Genetics Consortium

CRP – C-reactive protein

CVD – Cardiovascular disease

CVD-I – Olink Proteomics Cardiovascular I panel

CVD-II – Olink Proteomics Cardiovascular II panel

CVD-III – Olink Proteomics Cardiovascular II panel

DAMPs – Damage-associated molecular patterns

DBP – Diastolic blood pressure

DEG – Differentially-expressed gene

EA – European ancestry

EAF – Effect allele frequency

EAS – East Asian ancestry

EHR – Electronic heath record

eGFR – Estimated glomerular filtration rate

eQTL – Expression quantitative trait locus

FDR – False discovery rate

GRS – Genetic risk score

HDL – High-density lipoprotein

HIS – Hispanic ancestry

HRC – Haplotype Reference Consortium

IMP – Inosine monophosphate

IR – Insulin resistance
GLM – Generalised linear model
GMP – Guanine monophosphate
GO – Gene ontology
GRS – Genetic risk score
GUGC – Global Urate Genetics Consortium
GWAS – Genome-wide association study
INF – Olink Proteomics Inflammation panel
LD – Linkage disequilibrium
LDL – Low-density lipoprotein
LLOD – Lower limit of detection
MAC – Minor allele count
MAF – Minor allele frequency
MR – Mendelian randomisation
MSU – Monosodium urate
NAFLD – Non-alcoholic fatty liver disease
NO – Nitric oxide
NPX – Normalised protein expression
NSAID – Non-steroidal anti-inflammatory drug
OAT – Organic Anion Transporter-like
OR – Odds ratio
PD – Parkinson's Disease
PRPP – Phosphoribosylpyrophosphate
QC – Quality control
RGS – Reconstituted gene set
ROC – Receiver operating characteristic
ROS – Reactive oxygen species
SA – South Asian ancestry
SBP – Systolic blood pressure
SNP – Single nucleotide polymorphism
SUA – Serum uric acid
TG – Triglycerides
UA – Uric acid
VEP – Variant Effect Predictor
XO – Xanthine oxidase

IX. Note on genomic positions

All genomic positions throughout this thesis are reported with respect to the NCBI build 37 (hg19) version of the Human Genome.

Chapter 1 Introduction

1.1 Uric acid

1.1.1 What is uric acid?

Uric acid is a weak organic acid composed of carbon, hydrogen, oxygen and nitrogen (**Figure 1**). Clinical reference ranges vary between health providers, but 'normal' adult concentrations in serum generally range between 2.6 and 6.0 mg/dL in premenopausal women and 3.5 and 7.2 mg/dL in men and postmenopausal women. The discrepancy in serum urate levels between males and females begins at puberty and ends at menopause[1]. At physiological pH values, it is present in serum primarily in the form of urate ions. The solubility of uric acid in water is low, and in human serum the solubility limit is just 6.8 mg/dL. Above this, uric acid can precipitate out of solution as crystals of monosodium urate (MSU).

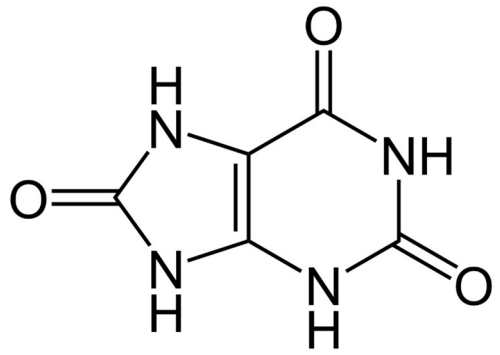


Figure 1 - Chemical structure of uric acid.

Uric acid is the end product of purine catabolism in humans, but far from being a simple waste product, its chemical properties suggest that it plays a complex and only partially understood role in a wide range of biological processes. Abnormally high or low levels of circulating uric acid (hyper- and hypo-uricemia) are both associated with a broad range of diseases. Most notably, hyperuricemia is linked to gout, a painful inflammatory arthritis triggered by monosodium urate crystals, as well as various cardiometabolic diseases, while hypouricaemia is associated with neurodegenerative disease. The importance of uric acid has been realised for

centuries, but the mechanisms linking it to human health and disease are still fields of active study.

The early history of uric acid is very much the history of gout, a disease known long before its primary cause. Descriptions of gout have been found as far back as ancient Egypt in 2640 BCE. Hippocrates' own aphorisms include such observations as *"Eunuchs do not take the gout", "A woman does not take the gout, unless her menses be stopped"* and *"In gouty affections, inflammation subsides within 40 days"*[2]. These early and remarkably accurate clinical observations reflect characteristics of gout still observed today. Gout has long been known as "the disease of kings" due it's occurrence mainly in those who could afford a decadent and luxurious lifestyle, to the extent that it has even been considered something of a status symbol in various periods of western history[2]. Gout is driven in large part by diet, and for much of history only the very wealthy could afford the purine-rich foods – meat, alcohol and seafood – that lead to hyperuricaemia. There were also observations that gout tended to run in families, the first recognition that gout – and urate homeostasis – could have a significant heritable component. There is even evidence that as early as the 6th Century CE colchicine was used as a treatment for gout[3] – a drug that is still used today as an anti-inflammatory agent for patients who cannot tolerate non-steroidal anti-inflammatory drugs (NSAIDs)

Though gout has been known for thousands of years, uric acid and the link between it and gout are comparatively recent discoveries. Uric acid was discovered in 1776 by Carl Wilhelm Scheele[4], a pharmaceutical chemist from Swedish Pomerania (modern day Germany) who was studying kidney stones (as well as discovering a number of organic acids, Scheele is noteworthy for being the first to make a number of chemical discoveries, including oxygen, hydrogen and chlorine, but being consistently beaten to publication, a misfortune that will still resonate with many modern scientists). Shortly after its discovery, an English chemist named Woolaston identified that urate was present within a tophus (a nodular growth characteristic of advanced gout) on his own ear[2]. By the end of the 19th Century, it was known that urate crystals injected into a joint could cause the formation of tophi, but conclusive scientific proof that the crystals seen in gout were monosodium urate did not come until the mid-20th century[5].

In the era of modern medicine, hyperuricaemia has been linked not only to gout, but to a wide range of diseases including type 2 diabetes[6], cardiovascular disease (CVD)[7], chronic kidney disease (CKD) and components of metabolic syndrome[8].

Hypouricaemia is associated with neurodegenerative diseases[9–12]. However, with the exception of gout, the mechanistic links behind these associations and the question of whether uric acid level has a causal role remain largely unresolved. The role of uric acid in disease thus remains an active and important area of medical research.

1.1.2 Urate production

Uric acid is the final oxidation product of purine metabolism in apes, including humans. Purines are heterocyclic aromatic organic molecules consisting of a pyrimidine ring and an imidazole ring – informally, they have a 6- and 5-sided double-ring structure containing nitrogen, as seen in the structure of uric acid itself in **Figure 1**. The group includes a wide range of molecules, including caffeine, but perhaps the most biologically significant are adenosine and guanine and their derivatives, together comprising one of the two groups of nucleotide bases.

Purines are metabolised primarily in the liver in a complex network of pathways summarised in **Figure 2**. In addition to exogenous purines obtained primarily through the diet, *de novo* synthesis of purines begins with the production of phosphoribosylpyrophosphate (PRPP) from adenosine triphosphate (ATP) and ribose 5′-phosphate, catalysed by PRPP synthetase – overactivity of this enzyme leads to overproduction of uric acid and gout[13]. PRPP is converted to inosine monophosphate (IMP), which can be converted to adenosine monophosphate (AMP) or guanine monophosphate (GMP).

AMP and GMP are removed by conversion back to IMP, which is dephosphorylated to inosine and converted to hypoxanthine. Xanthine oxidase (XO) catalyses the oxidation of hypoxanthine to xanthine, and xanthine to urate. Adenosine can also be deaminated directly to inosine, and guanine converted to xanthine.

As **Figure 2** shows, urate production is coupled to multiple different pathways including glucose and fructose metabolism, and any process consuming or producing ATP will alter these pathways - urate production rate is actually in index for ATP-synthesis in hepatocytes[14]. This diagram is by no means exhaustive, but it is already clear from this limited view how increased alcohol consumption, diabetes, high fructose intake and purine-rich diets can all lead to increased serum urate levels.

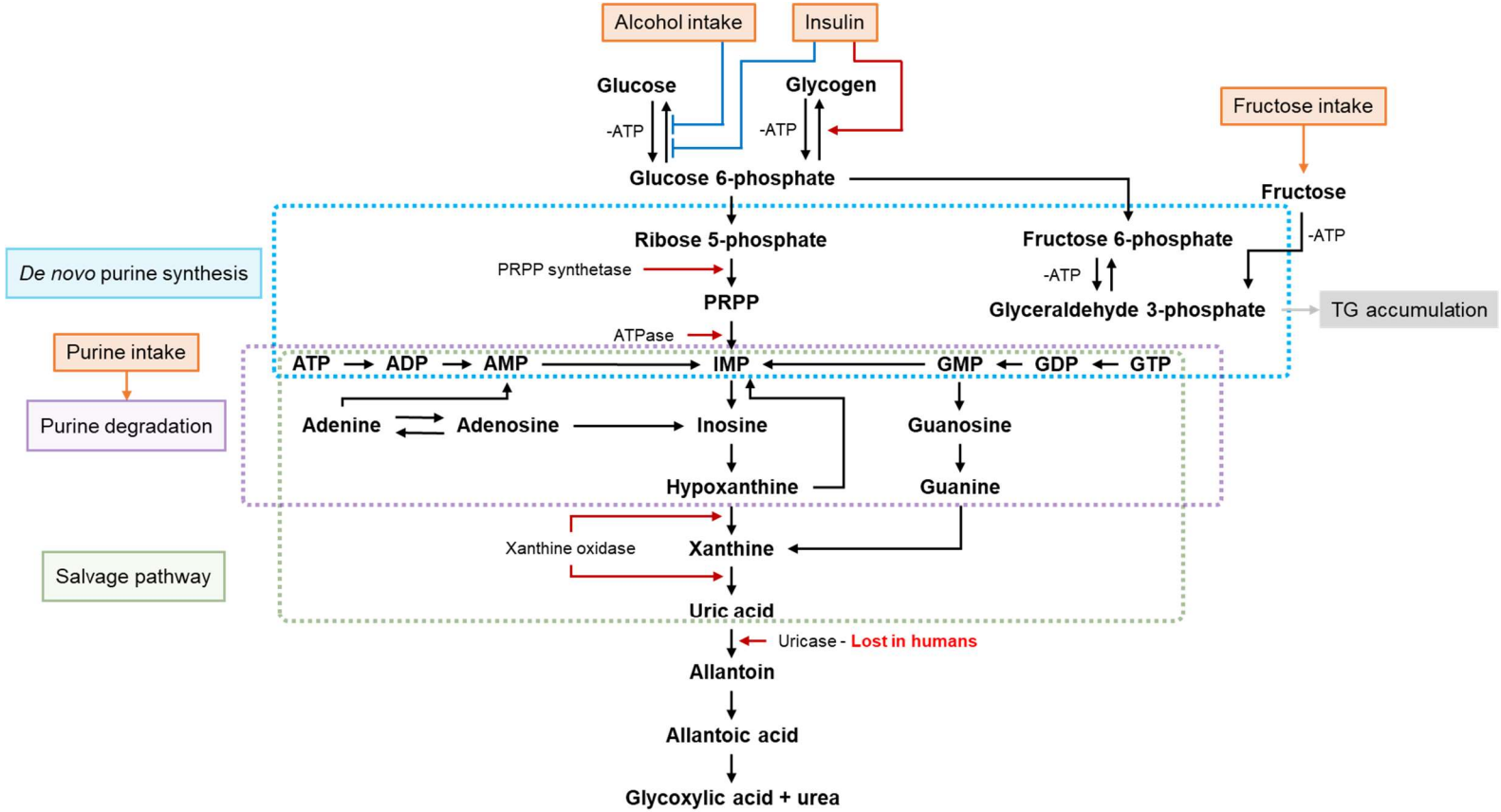


Figure 2 - Uric acid pathways.

Figure modified from Kushiyama *et al.* (2016)[15]. Abbreviations: (A/G)(TP/DP/MP) – adenosine/guanine tri/di/monophosphate. IMP – inosine monophosphate. TG – triglyceride. PRPP – phosphoribosylpyrophosphate.

1.1.2.1 The loss of uricase in humans

In all mammals, including humans, uric acid is produced from xanthine, catalysed by the xanthine oxidase enzyme. In most mammals, uricase (also called urate oxidase) further oxidises urate to 5-hydroxyisourate, which is converted by two further enzymes into allantoin, a highly-soluble waste product that is easily excreted in urine[16] (this is simplified into a single step in **Figure 2**). However, the apes, the group which includes humans, have lost uricase activity, breaking the pathway and leaving urate as the final end product of purine oxidation. Consequently, physiological levels of urate in serum are an order of magnitude higher in humans than most mammals (0.5 - 0.8 mg/dL in mice), and they vary over a much wider range.

An interesting exception is seen in Dalmatian dogs, which have uric acid levels as much as ten times higher than mongrel dogs[17,18], giving them a predisposition to urinary calculi formed from urate. This is not caused by mutations in uricase[19], instead appearing to be due to a defect in the *SLC2A9* gene, which encodes a urate transporter. This leads to reduced urate transport into the liver, the site of uricase activity, and reduced excretion in the kidney[20]. This mutation was likely fixed in the breed by artificial selection for their distinctive spotted appearance.

By contrast, the loss of uricase activity in humans may in fact have had a selective advantage. Loss of function appears to have been gradual, caused by the accumulation of several mutations over millions of years[21,22], and has occurred independently in several different lineages of higher primates. This suggests that increased concentrations of uric acid in the blood were beneficial. Various possible reasons for this have been suggested. The loss of uricase may have compensated for the earlier loss of ascorbic acid production in humans, as uric acid can also act as an antioxidant, and accounts for roughly half the antioxidant capacity of plasma[23]. It has also been suggested to be useful in times of starvation, as uric acid has been linked to the regulation of hepatic glucose production and the timing of the loss of uricase activity appears to coincide with a period of climatic change[24]. Kratzer *et al.* suggest that the progressive loss of uricase may have allowed the accumulation of fat via fructose metabolism, a potential advantage when climate shifts at the end of the Oligocene lead to the replacement of rainforests in Europe and Asia with less energy-dense temperate forest[25]. Regardless of the selective driver behind the loss of function, the accumulation of uric acid has a wide range of physiological and ultimately clinical consequences.

1.1.3 Properties of uric acid

Uric acid is a molecule with a complex range of biologically relevant and sometimes seemingly contradictory properties that make it a particularly interesting molecule to study. It is perhaps most famously known as an antioxidant, and indeed *in vitro* experiments in the 1980s showed that it is a powerful scavenger of damaging oxygen, peroxyl and hydroxy radicals, forming the basis of the theory that uricase activity was lost to compensate for the loss of ascorbate metabolism[23]. The antioxidant properties of uric acid have been proposed as the reason why gout appears to be protective against neurodegenerative diseases caused by radical damage[26–28]. Urate is a potent scavenger of peroxynitrite ($ONOO^-$), a highly reactive oxidant and nitrating agent which can induce protein nitrosation and lipid and protein oxidation, as well as causing apoptosis through general oxidative damage to DNA and cellular components (**Figure 3a**). Uric acid neutralises peroxynitrite many times faster than abscorbate[29]. In mice with experimental autoimmune encephalomyelitis (an animal model of demyelinating brain disease), uric acid was shown to have strong dose-dependent therapeutic effects, including blocking the peroxynitrate-mediated nitrosation of neuronal proteins[28].

However, it has since been observed that urate cannot scavenge all radicals, and does not consistently act as an antioxidant, requiring specific conditions to do so. While urate can inhibit Cu^{2+}-induced oxidation of low-density lipoprotein (LDL) cholesterol (**Figure 3b**), in the presence of transition metals it can increase the oxidation of oxidised LDLs. In plasma, urate acts as an anti-oxidant, but in intracellular conditions, it can be a prooxidant. Urate can form radicals when reacting with other oxidants, particularly lipids. Additionally, the hydrophobic environment created by lipids limits the antioxidant capacity of uric acid, which has been proposed as an additional complication of the elevated serum urate levels seen in obesity[30]. *In vitro* experiments in rats have identified a role for uric acid in increasing monocyte chemoattractant protein-1 (MCP-1, also known as CCL2) production, a protein involved in vascular inflammation, via a mechanism that is driven by oxidative activity[31] (**Figure 3c**).

Uric acid also interacts with nitric oxide (NO), a gas originally identified as an endothelial cell-derived relaxing factor, but which has emerged to be an important regulator of cardiovascular function linked to hypertension, cardiovascular disease and metabolic syndrome. NO is a vasodilator, and reduced levels can lead to

endothelial dysfunction, a pathogenic event in which excess constriction of the blood vessels leads to inflammation, platelet activation and increased permeability to lipoproteins and other toxins. Endothelial dysfunction often precedes more advanced cardiometabolic disease. Oxidative stress is a primary driver of endothelial dysfunction, which would suggest a protective role for uric acid, but hyperuricaemia was paradoxically found to induce it[32]. This effect was found to be driven by urate directly interacting with NO to form 6-aminouracil, which depletes NO, a reaction which could be blocked by another antioxidant, glutathione. When urate levels exceed the capacity of glutathione to supress this depletion, as in hyperuricaemia, endothelial dysfunction ensues[33] (**Figure 3c**).

To further complicate the paradoxical role of uric acid, the xanthine oxidase pathway, which produces uric acid, is actually one of the two major systems driving vascular oxidative stress. The process of converting xanthine to uric acid releases reactive oxygen species (ROS), highly oxidative molecules which cause endothelial dysfunction. For this reason, the XO inhibitor allopurinol, usually prescribed to treat gout, has been proposed as a treatment for hypertension. The other major process driving oxidative stress is NADPH oxidation, which uric acid can stimulate in adipocytes[34], **Figure 3c**.

The crystallised form of uric acid also plays a role as a danger signal. MSU crystals have been shown to trigger immune and inflammatory responses in various ways depending on the microenvironment in which the crystallisation occurs[35]. Uric acid has also been identified as a damage-associated molecular patterns (DAMPs) component, being released from damaged cells, particularly after necrotic cell death, acting as an attractant for leukocytes and mesenchymal stromal cells, promoting inflammation and repair[36].

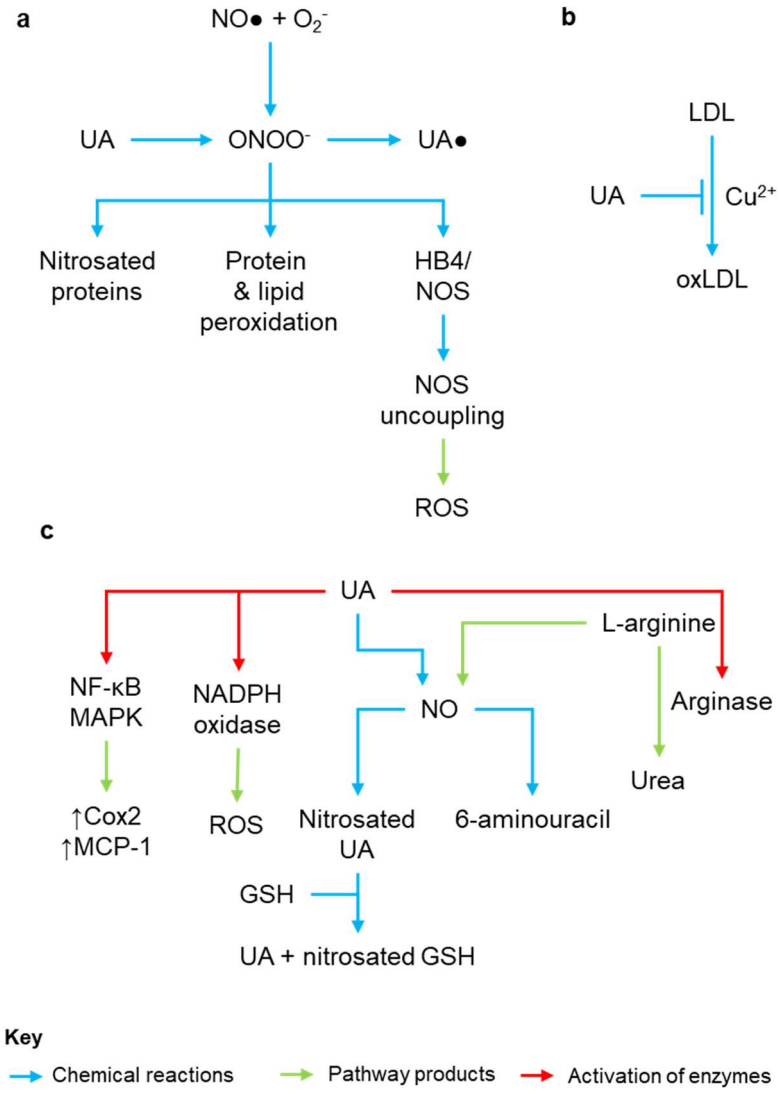


Figure 3 - Antioxidant and prooxidant effects of uric acid.

Figure adapted from So and Thorens (2010)[37]. **a** Peroxynitrite, produced by nitric oxide (NO●) radicals reacting with superoxide (O_2^-), can cause damage by nitrosating proteins, oxidising proteins and lipids. It also blocks tetrahydrobiopterin (HB4), an important cofactor necessary for nitric oxide synthase (NOS), which leads to production of reactive oxygen species (ROS). Uric acid (UA) can inactivate $ONOO^-$, producing uric acid radicals (UA●) which are in turn neutralised by ascorbic acid. **b** UA can prevent copper ion (Cu^{2+}) mediated oxidation of LDL. **c** In vascular smooth muscle cells, UA activates the NF-κB and MAPK pathways, increasing cyclooxygenase and MCP-1 production. In adipocytes, UA uptake activates NADPH oxidase, leading to ROS production. UA enhances arginase activity, which reduces NO levels by diverting L-arginine, necessary for NO production, into urea production. UA can react with NO to produce nitrosated UA, which can transfer its nitroso group to glutathione (GSH). In the presence of oxygen, this reaction produces stable 6-aminouracil, depleting NO.

1.1.4 Uric acid transport

Serum urate levels are the result of a balance between its production, primarily in the liver, and its excretion in the kidney and the gut. Transport proteins play a key role in regulating this balance, and consequently many have been identified as modulating serum urate levels and gout, both through genome-wide association studies (GWAS) and functional work.

The kidney is a key organ in urate homeostasis, accounting for an estimated 70% of urate excretion[37]. Urate is filtered in the glomerulus and a combination of reabsorption and secretion is orchestrated in the proximal convoluted tubule by various transporters, summarised in **Figure 4**. A net reabsorption occurs such that the fractional excretion of urate (the fraction of the total filtrate which is ultimately excreted) is normally ~6-8%.

Among the transporters are several proteins from the organic anion transporter-like (OAT) family, encoded for by genes in the *SLC22A* family. URAT1 (encoded by *SLC22A12*) is located on the apical membrane of proximal tubule epithelial cells, and transports urate into the cell in exchange for chlorine ions or organic anions[38]. Antiuricosuric agents (uricosuria is the condition of having uric acid in the urine, thus antiuricosuric drugs promote reuptake of urate) such as lactate, pyrazinoate and nicotinate can serve as substrates for URAT1 and thus increase urate uptake, while benzbromarone, probenecid and losartan, all used to treat gout, inhibit it and increase renal uric acid clearance[38]. Inactivating mutations in this protein have been found to cause idiopathic renal hypouricaemia[38].

OAT4 (*SLC22A11*) and OAT10 (*SLC22A13*) are also expressed at the apical membrane and thought to similarly reabsorb urate from the tubule lumen[39–41]. OAT4 can transport urate into cells in exchange for dicarboxylates[39]. OAT10 is a high-affinity nicotinate transporter[41]. OAT1 (*SLC22A6*) and OAT3 (*SLC22A8*) also function as anion/dicarboxylate exchangers and can transport urate into cells but are expressed at the basolateral membrane. They have been suggested to play a role in urate secretion, rather than uptake, based on the predicted gradient of dicarboxylate across cell membranes[42–45].

MRP4 (*ABCC4*) and ABCG2 (*ABCG2*) are both expressed on the apical membrane, but export urate out into the lumen in an ATP-dependent manner, therefore driving urate secretion[46–48].ABCG2 is the major urate transporter in the gut[49], and has been

suggested to play a compensatory role in individuals with CKD where renal urate excretion is impaired[50]. A Q141K mutation in this gene (rs2231142) confers a greatly increased risk of gout[51].

Sodium phosphate transporter 4 (NPT4, *SLC17A3*) is a voltage-driven transporter expressed on the apical membrane that secretes urate[52]. NPT1 (*SLC17A1*) is similarly expressed on the apical membrane and a hypermorphic mutation in the protein has been linked to reduced risk of renal underexcretion gout[53].

Currently, only one transporter is known that transports reabsorbed urate into the interstitium for circulatory system uptake – GLUT9, encoded by *SLC2A9*. This protein is unusual in that its primary function was discovered following GWAS – prior to this it was known as a glucose/fructose transporter of very low activity[54]. Early GWAS of serum urate levels identified a strong association signal in *SLC2A9* explaining around 3.5% percent of variance in serum urate levels, with various experimental methods confirming its role as a urate transporter with low affinity but high capacity[55–57]. Two alternatively-spliced variants exist, GLUT9a and GLUT9b. In humans, GLUT9b is only expressed in the kidney and placenta, while GLUT9a is more broadly expressed in kidney, liver, intestine, leukocytes and chondrocytes[58]. The longer isoform GLUT9a traffics to the basolateral membrane of the kidney proximal tubule cells and plays a major role in urate reuptake, GLUT9b locates to the apical membrane. Its transport functions are dependent on membrane potential, but are not sensitive to Na^+ or Cl^- concentrations[59]. Loss of function mutations in GLUT9 cause monogenic hypouricaemia, and show more severe phenotypes than loss of function mutations in URAT1[60].

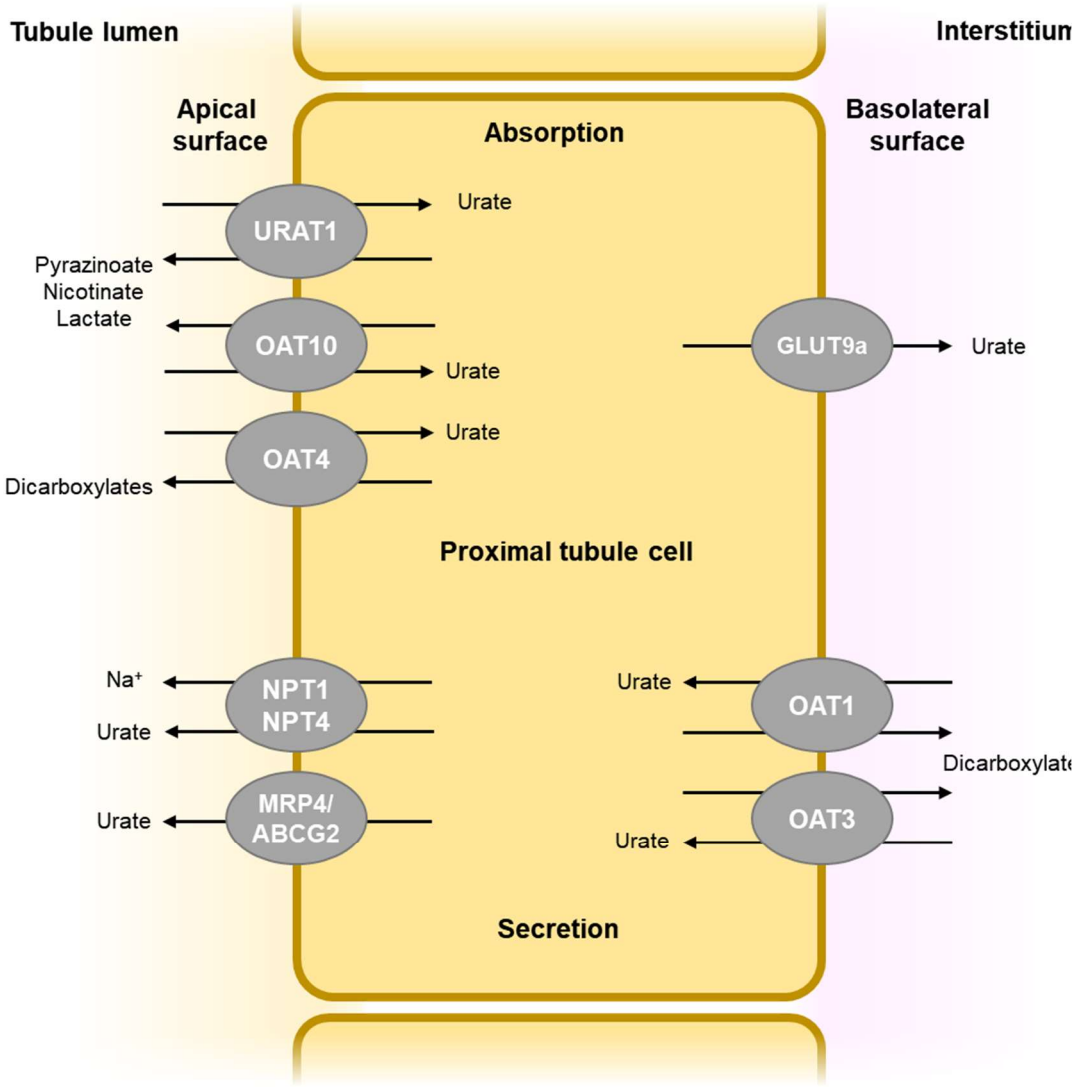


Figure 4 - Uric acid transporters in renal proximal epithelial cells.
Figure adapted from So and Thorens (2010)[37].

1.1.5 Effect of sex hormones

As mentioned above, serum urate levels are significantly higher in men than in women, and postmenopausal women have higher levels than premenopausal. It has been proposed that hormonal changes explain some of these differences – oestrogen appears to increase renal clearance of uric acid[61–63], even when administered to males[62], but the mechanism remains unknown.

Male-to-female transsexuals showed reduced uric acid after hormone therapy, but while female-to-males showed increased uric acid, it remains unclear whether

testosterone also plays a role[64]. A recent study of postmenopausal women found that uric acid was lowered in a group receiving combined oestrogen and progesterone therapy, but no significant difference in a group receiving only oestrogen. This could suggest the regulatory system is more complex, but could also reflect the smaller sample size in the oestrogen-only group[65].

1.1.6 Genome-wide association studies of serum urate

GWAS have had a major impact on medical research over the last decade, and uric acid is no exception. The principles behind GWAS will be explored in more detail in Chapter 3, but to briefly summarise, the aim of a genome wide association scan is to identify genetic variants which are associated with a trait of interest – whether a disease or a quantitative trait, such as serum urate levels. Advances in computational power, recruitment of increasingly large research cohorts and development of sophisticated statistical techniques to maximise power have led to a steady increase in the number of regions of the genome known to have a link with serum urate levels, even though the mechanism behind the association is not always clear.

1.1.6.1 *SLC2A9*/GLUT9

The earliest GWAS of serum urate identified *SLC2A9* as being significantly associated with serum urate[55–57,66]. The association between variants in this gene and levels of serum urate is so strong that it was detectable in fewer than 1,000 individuals. Prior to these associations, the encoded protein GLUT9 was suspected by structural homology to be a glucose or fructose transporter, and while it does show transport ability for these molecules, functional experiments in *Xenopus* oocytes demonstrated that it had a much greater urate transporting capacity[55]. As previously mentioned, this was a remarkable example of GWAS leading directly to a novel functional insight – most GWAS variants identify regulatory regions of the genome that are not so immediately interpretable.

Strong sex-specific effects were also remarkable at this locus, with the same variant explaining 1.2% of the variation in serum urate levels in men, but 6% in women[55,56]. An early study of four directly-genotyped polymorphisms within the *SLC2A9* locus within two cohorts, one a population sample of 800 and the other a case-control sample with 1,038 severely obese (body mass index (BMI) between 33 and 92 kg/m^2) and 831 controls (BMI < 33kg/m^2), suggested that obesity (body mass index (BMI) > 40kg/m^2) increased effects of these variants, with a stronger modulating effect in men

than women[67]. A later study in a predominantly female sample, with 520 obese (BMI > 35kg/m^2), and 540 controls (BMI < 25kg/m^2), reported the opposite, stronger associations in lean than obese[68], but the sex differences and different criteria for controls make it difficult to draw direct comparisons.

A more recent GWAS with a somewhat larger sample size stratified into lean, overweight and obese individuals did not identify statistically significant differences between BMI strata at this locus[69].

1.1.6.2 Early meta-analysis of serum urate

As the GWAS field matured, meta-analysis of multiple cohorts allowed increased statistical power. Power to detect variants increases with sample size, effect size and minor allele frequency, so larger sample sizes were necessary to detect the effects of rarer variants of similar effect size to those discovered in early studies, or variants with smaller effects.

An early meta-analysis by Dehghan *et al.* (2008) identified variants in the *ABCG2* and *SLC17A4-SLC17A1-SLC17A3* loci[51]. The Q141K variant referred to above was identified in this analysis, which later analyses confirmed to be a strong risk variant for gout commonly segregatng in populations of European ancestry. Sex-specific effects were also identified in *ABCG2*, with stronger effect in men. Prior to this, ABCG2 was known as an efflux transporter for multiple drugs, but also purine nucleoside analogues, making it a promising candidate for urate transport. Follow up experiments confirmed its urate transporting activity[70]. *SLC17A1* and *SLC17A3* encode the urate transporters NPT1 and NPT4 respectively, though at the time only NPT1 had been demonstrated to have a urate transporting effect in model systems[71]. Dehghan *et al.* also constructed a genetic risk score using variants in these two genes and *SLC2A9* to test for association with gout, remarkably identifying a 41-fold difference in risk between the lowest and highest risk groups.

Larger meta-analyses identified more signals. Kolz *et al.* (2009) meta-analysed data from 14 studies comprising over 28,000 individuals. They identified six additional loci, including two close the *SLC22A11* and *SLC22A12* genes. *SLC22A12* encodes URAT1, the very first (and thoroughly) characterised urate transporter, identified before the GWAS era by sequence homology with known OAT proteins[38], while *SLC22A11* encodes OAT4. The study also identified monocarboxylic acid transporter 9 (MCT9, *SLC16A9*), glucokinase regulatory protein (*GCKR*), Carmil (*LRRC16A*), and near PDZ domain-containing 1 (*PDZK1*).

OAT4 is a urate transporter active in the kidney, as described above.

MCT9 transports monocarboxylic acid across cell membranes and is expressed in the kidney. Its role in urate homeostasis remains unknown, but later work identified a K258T mutation that is associated with renal overload gout (gout driven by overproduction of urate, rather than underexcretion)[72] and Kolz et al noted that variants associated with urate levels at this locus also influenced carnitine levels suggesting a link to energy production and fatty acid metabolism.

Carmil is a regulator of actin polymerisation, and an abundant protein in the kidney and epithelia.

GKCR encodes glucokinase regulatory protein, a regulator of glucokinase, the enzyme which phosphorylates glucose, a key first step in *de novo* synthesis of uric acid (see **Figure 2**). The identified variant is also associated with a range of metabolism- and diabetes-related traits.

PDZK1 encodes a scaffolding protein thought to interact with and regulate OAT4, URAT1 and NTP1 (*SLC17A1*) via N-terminal PDZ motifs, and has been proposed to link URAT1 and NPT1 into a functional complex that both secretes and reabsorbs urate, allowing precise regulation[73].

1.1.6.3 Global Urate Genetics Consortium

The largest published GWAS of serum urate to date was run by the Global Urate Genetics Consortium (GUGC) and published in Köttgen *et al.* (2013)[74]. This GWAS identified 18 new signals, from a total of 28, with novel loci reported at *TRIM46*, *INHBB*, *SFMBT1*, *TMEM171*, *VEGFA*, *BAZ1B*, *PRKAG2*, *STC1*, *HNF4G*, *A1CF*, *ATXN2*, *UBE2Q2*, *IGF1R*, *NFAT5*, *MAF*, *HLF*, *ACVR1B-ACVRL1* and *B3GNT4*. This analysis reported 7.0% of variance in serum urate levels was explained by these 28 single nucleotide polymorphisms (SNPs) (5.2% by the previously known loci and 3.4% by *SLC2A9* and *ABCG2* alone). Heritability of serum urate concentration is estimated at 40-70%, suggesting that there are more genetic variants to be found.

The novel regions identified in this meta-analysis were within or close to genes without an obvious role in urate transport. Expression quantitative trait locus (eQTL) and network analysis suggested that the implicated genes in the novel loci play roles in glucose and lipid metabolism. Inhibins-activins signalling pathways were suggested by protein-protein interaction networks, but it remains unclear which biological functions these contribute to.

1.1.6.4 BioBank Japan

The latest development in the GWAS field is the rise of the 'supercohort', single cohorts of very large size – in excess of a hundred thousand individuals, with some projects aiming for a million. These are often national prestige projects, such as the UK Biobank[75,76], a project notable for making its data available to any researcher subject to project approval. UK Biobank has not yet released biochemistry data (this is discussed in more detail in Sections 4.4.8 and 5.1.1), but the BioBank Japan project has recently published GWAS results for 58 quantitative traits, including uric acid[77].

Though the sample size of the analysis is slightly smaller than the GUGC meta-analysis (>109,000), ten of the twenty-seven loci they reported were novel. This will be in part due to the different ancestry of the cohort – most GWAS research has been on European-ancestry populations – but will also reflect improvements in genotype imputation, as denser and more accurate reference panels have been released since the GUGC analysis (see Chapter 3). Novel loci were identified at *USP34*, PRDM8/FGF5, *UNCX/MICALL2*, *TP53INP1/NDUFAF6*, *BICC1*, *FAM35A*, *EMX2/RAB11FIP2*, *SBF2*, *MPPED2/DCDC5* and *LOC101927932*. Little interpretation of these loci was made in the paper, which is an emerging characteristic of 'supercohort' publications: associations with dozens of phenotypes are published but it is left to the reader to interpret their relevance. In part this is an inevitable consequence of publishing such broad research, but it is also symptomatic of the increasingly complex associations revealed by GWAS analysis. Interpreting individual results is difficult, and network- and pathway-based approaches with functional validation will become indispensable in the future.

1.2 Clinical and epidemiological relevance

The ancestral loss of uricase may have conferred a selective advantage, but it has come at a cost to the health of modern humans. While mechanisms to regulate serum urate levels are sophisticated, modern western diets are purine-rich, containing large amounts of alcohol, red meat and seafood. This increases the pool of exogenous purines and consequently leads to elevated serum urate levels.

Cross-sectional studies have found variation of serum uric acid level with age is complex and varies between sexes, but are inconsistent in their conclusions, likely because they provide a snapshot of the population rather than following the same

individuals over time[78–80]. In a longitudinal study of 80,506 Japanese office workers, serum urate levels were found to increase with age independently of changes in BMI and alcohol consumption[81]. This was particularly strong in women between 40 and 70, corresponding to the menopause.

Although uric acid has been associated with a wide range of diseases, evidence that it is a risk factor rather than a consequence is rarely consistent, except in the case of gout. A recent comprehensive umbrella review compared evidence from 101 publications, encompassing systematic reviews, meta-analyses of observational studies, meta-analyses of randomised controlled trials and 107 Mendelian randomisation (MR) studies, exploring a total of 136 unique health outcomes[82]. Evidence was classified as 'convincing', 'strongly suggestive', 'suggestive' or 'weak' depending on criteria including P-values, sample size and heterogeneity. No observational studies were found to have 'convincing' evidence, but 'strongly suggestive' evidence was found for associations between serum urate and risk of heart failure, hypertension, impaired fasting glucose or diabetes, CKD and coronary heart disease (CHD) mortality. Four outcomes had significant P-values for MR studies – diabetic macrovascular disease, arterial stiffness, renal events and gout, but, of these, only gout had convincing evidence. Recurrence of nephrolithiasis (kidney stones) showed some evidence of association in randomised controlled trials, but not in observational studies or MR.

However, this review is not without limitations. As Borghi points out in his response to the article[83], serum uric acid levels may not be a good discriminatory tool, as the same level can arise through different mechanisms – overproduction or underexcretion, through a variety of pathways – and thus may not be expected to associated similarly with disease. Additionally, he suggests that the role of xanthine oxidase and oxidative stress cannot be dismissed, and that the considerable heterogeneity in thresholds used for hyperuricaemia between studies causes difficulties in interpretation. This review is cited here not as proof that abnormal serum urate levels are or are not causal for a given disease, but rather as a demonstration of how complex the question remains despite so many attempts to address it.

1.2.1 Gout

The dominant cause of hyperuricaemia in individuals with gout appears to be underexcretion of urate, rather than overproduction – when compared to healthy

controls, gout patients showed lower uric acid clearance (filtration from serum), fractional excretion of uric acid (percentage of uric acid filtered by the kidney that is then excreted in the urine) and urinary uric acid to creatinine ratio than controls[84]. While allantoin is highly soluble in water under physiological conditions, allowing it to be easily eliminated in urine, uric acid is considerably less so. The solubility limit of 6.8mg/dL is within the normal range of serum urate concentrations in humans, and as a result, spontaneous formation of MSU crystals can occur. These can grow into nodular masses called tophi in advanced gout, usually around ten years after the initial flares. This most often occurs in the synovial fluid of the joints, but tophi can be found on the elbows, upper ear cartilage and joint surfaces. However, not all hyperuricaemic individuals exhibit MSU crystals. The factors controlling crystal formation remain poorly understood, but factors affecting the solubility of urate such as temperature and pH may contribute.

Gout is characterised by painful 'attacks' or 'flares', periods of intense but self-limiting pain. These are triggered by macrophages interacting with MSU crystals and activating the NLRP3 inflammasome[85], releasing interleukin 1β and activating neutrophils and mast cells, which release a host of pro-inflammatory cytokines, chemokines and ROS, amplifying the response[86]. However, only around 20%of hyperuricaemic individuals experience gout attacks[87]. The reasons for this are unclear, though perhaps related to variation in the concentration at which crystallisation occurs. It remains an area of active research which may lead to new therapeutic options for gout sufferers.

The prevalence of gout is increasing worldwide, but it is asymmetrically distributed, being higher in developed countries and particularly high in Pacific island populations[88]. In a study in the US, prevalence increased by two cases per 1,000 over ten years, but in over 75s the rate doubled from 21 per 1,000 in 1990 to 41 per 1,000 in 1999[89]. Prevalence was estimated at around 4% in the US in 2008[90]. Gout is also on the rise in the UK, from 1.42% in 1997 to 2.49% in 2012[91]. Prevalence is much higher among men (3-6%) than women (1-2%), and but increases in women after menopause[88], reflecting similar patterns in serum urate levels.

1.2.2 Cardiometabolic disease

Hyperuricaemia frequently occurs in patients with metabolic syndrome – a collective term for a cluster of associated symptoms that put patients at increased risk of CHD,

stroke and diabetes. Its definition has varied over time – hyperuricaemia has previously been one of the defining components – but it broadly reflects central obesity, hypertension, hyperglycaemia, and dyslipidaemia. Obesity – visceral adiposity in particular – is closely correlated to hyperuricaemia[92].

Metabolic syndrome is driven by a complex and highly interconnected set of pathways (**Figure 5**). It is characterised by insulin resistance and high levels of free fatty acids, which can be due to a combination of genetics and lifestyle and are often associated with obesity. Adipose tissue releases inflammatory factors, angiotensin and adipokines, which suppress insulin signalling[93,94]. Angiotensin release leads to hypertension, exacerbated by increased sodium retention in the kidney driven by high levels of insulin, a consequence of insulin resistance. In the liver, free fatty acids accumulate from insulin-resistant adipose tissue, altering the balance of LDL and high-density lipoprotein (HDL) cholesterol and increasing glucose production. Combined with reduced glucose uptake by muscle and adipose tissue, this leads to hyperglycaemia[95].

Insulin drives increased reuptake of urate in the kidney – *in vitro* experiments have shown that insulin significantly increases endogenous levels of URAT1[96]. This is one pathway by which metabolic syndrome is connected to hyperuricaemia. Altered glucose metabolism also increases uric acid production, through ATP depletion and reduced phosphate levels. Inflammation and oxidative stress increase and are increased by elevated uric acid levels, leading to reduced bioavailability of NO and subsequent endothelial dysfunction.

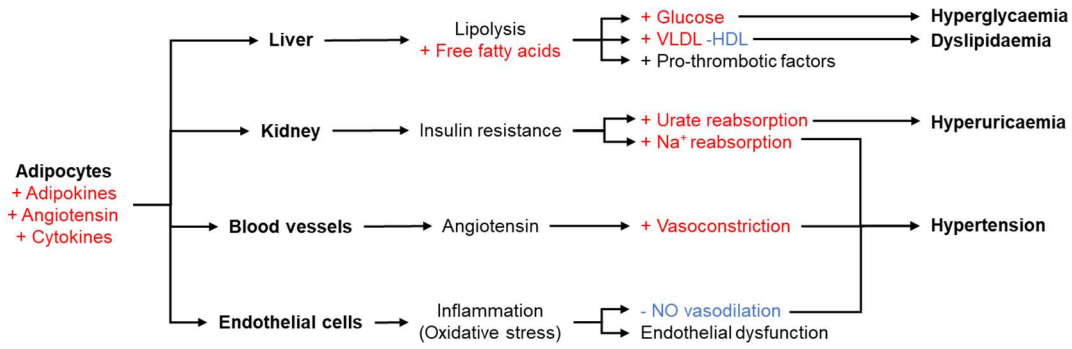


Figure 5 – Mechanisms of metabolic syndrome.

Figure adapted from Battelli *et al.* (2018)[95]. NO: nitric oxide. VLDL: Very low-density lipoprotein. HDL: high-density lipoprotein. Na^+: sodium ions.

Hyperuricaemia is commonly found in CVD patients, but whether it is a risk factor or a consequence remains unclear. Mendelian randomisation analyses have found evidence both for and against causality[82]. Early work in the Framingham Heart Study concluded that hyperuricaemia was a consequence of known cardiovascular risk factors[78] but several studies have suggested that uric acid plays a direct role in endothelial dysfunction[97,98], which can lead to CVD. There is evidence that urate can increase the oxidation of already oxidised LDL cholesterol, leading to increased risk of thrombosis and atherosclerosis. The production of ROS involved in urate synthesis and the inflammatory response it triggers can also lead to endothelial damage (for details see Section 1.1.3). The direct effect of uric acid on endothelial function in vivo has been difficult to investigate due to confounding by comorbidities, reverse causations and canalization. In non-hypertensive hyperuricemic adult individuals, the XO inhibitor allopurinol, which lowers serum urate by preventing xanthine oxidation, and the linked ROS production, was shown not to affect endothelial function[99]. This suggested that serum urate level *per se* may not be directly implicated in regulating endothelial dysfunction.

Hypertension has been suggested to passively increase serum urate levels, as early hypertension leads to decreased renal blood flow, which could alter the rate of urate secretion[100]. However, there is evidence that hyperuricaemia may play a direct role in hypertension through its activation of the renin-angiotensin system and its direct effects on NO (see Section 1.1.3) and trials have shown that allopurinol can reduce

blood pressure in adolescents[101]. There is also evidence that in children hyperuricaemia is associated with primary hypertension, but not secondary, suggesting it plays a causal role[102]. Mendelian randomisation studies remain inconclusive[82], perhaps limited by current knowledge of genetic associations, and the pleiotropic effects of the different associated variants.

A post-hoc analysis of the EURIKA study [103], a cross-sectional study of patients with no clinical disease but at least one risk factor for cardiovascular disease, identified an increased risk of death in patients with higher serum urate levels. This effect remains significant after accounting for diuretic use or impaired renal clearance. As observed by Carluccio *et al.* [104], this suggests that the risk is associated with increased urate production by XO, rather than accumulation of serum urate *per se* , which could also occur due to reduced excretion.

The ready availability of urate-lowering drugs means that a better understanding of the relationship between cardiovascular disease and uric acid could lead to new therapeutic options without new drug development, if the relationship could be understood.

Hyperuricaemia frequently co-occurs with type 2 diabetes, but again, the causality is unclear. Fructose, a major dietary cause of diabetes, is rapidly converted to fructose-1-phosphate, which results in a fall of intracellular phosphate and ATP, stimulating the direct generation of uric acid though nucleotide turnover[6]. Uric acid clearance is inversely correlated with insulin resistance (IR).

1.2.3 Renal disease

Serum uric acid levels are inversely correlated with measures of glomerular filtration rate (a measure of kidney function) and approximately 20 to 60% of gout patients also have some renal dysfunction[105]. Kidney disease will inevitably lead to abnormal concentrations of many substances in the blood, but there it has been suggested for decades that hyperuricaemia is more than just a symptom. A study published in 1978 of patients with normal kidney function comparing hyperuricaemic (>8.0 mg/dL) to normal (<5.0 mg/dL) patients found nearly a threefold increase in risk of developing renal deficiency within two years for men, and a tenfold risk for women[106]. More recently, studies in rat models have found that hyperuricaemia-induced kidney damage can be reversed with urate-lowering agents[107]. A recent meta-analysis found elevated uric acid levels to be predictive of new-onset CKD[108], and another cautiously

suggested that urate-lowering medication may slow the progression of CKD, but larger studies were required to test this assumption[109].

It has been suggested that activation of the renin-angiotensin system by uric acid is one mechanism by which uric acid could lead to kidney disease, as this raises systemic and glomerular pressure, which can cause fibrosis of renal cells[110].

1.2.4 Neurodegenerative disease

Low serum uric acid levels are also associated with neurogenerative disease – associations have also been found between lower levels of uric acid and multiple sclerosis[9], Parkinson's disease[10], Huntington's disease[111], Alzheimer's disease[112], and optic neuritis[113], though these are all small studies.

These associations may be consequences of diseases, but could also be due to uric acid's role as an antioxidant, with its $ONOO^-$ scavenging capacity suggested as a particularly important mechanism[100]. $ONOO^-$ and similar radicals are thought to be involved in the inflammation, demyelination and oxidative stress involved in various forms of neurodegenerative disease.

1.3 Project Aims

It is increasingly apparent that, far from the inert waste product it was once assumed to be, uric acid is a biologically active molecule that plays a range of roles in inflammation, signalling and protection against oxidative stress. As both hyperuricaemia and hypouricaemia are associated with disease, with some biologically plausible mechanisms for their pathogenicity, a better understanding of the pathways regulating uric acid levels could provide novel insights into these associated diseases. In addition to providing biological insight, this greater knowledge will be essential for guiding the assessment of the therapeutic potential of urate-altering drugs.

This project aims to further investigate serum urate associations and regulation, with a particular emphasis on exploring their genetic underpinning. It is hoped that this will help define specific pathways altering serum urate levels and lead to better understanding of the associated diseases and conditions.

This will take two primary approaches. First, high-dimensional molecular phenotyping data, I aim to identify proteins and lipids which are associated with serum urate levels.

The significantly associated traits and the look up for shared genetic factors which could help interpret associations are detailed in Chapter 2.

Second, I will use genetic association methods to identify new regions of the genome influencing serum urate levels, first using data from our own cohorts, detailed in Chapter 3 and then as part of a large GWAS consortium, CKDGen, detailed in Chapter 4.

Finally, in Chapter 5, I will detail the current status of a work-in-progress collaboration with Professor Tony Merriman (University of Otago) to identify genetic variants which are protective against gout.

Taken together, the work in this thesis represents a body of new knowledge on the regulation of serum urate and its interactions with other phenotypes in health and disease.

Chapter 2 Phenotype Correlations

2.1 Background

Uric acid synthesis is tightly linked to ATP homeostasis and metabolic pathways, and accordingly serum urate levels are correlated with a variety of phenotypes. Some have been known for many years, such as BMI, HDL cholesterol levels or alcohol consumption[114]. While the associations often seem intuitive, the underlying mechanisms are partly or wholly unclear, as these complex phenotypes are the compound result of multiple different physiological pathways, governed by the complex interplay of both genetic and environmental factors.

Less complex endophenotypes are an attractive way to attempt to untangle the relationships between uric acid and high-level phenotypes. As they are closer to the gene level, the relationship between genotype and phenotype is generally less polygenic and the underlying mechanism can be easier to interpret. This has become a particularly viable method with the advent of relatively cheap, high-throughput omic phenotyping platforms.

'Omics' is a somewhat loosely defined collective term for fields that focus on the characterisation and quantification of a large number of related biological measurements – strictly speaking, the word encompasses all fields in biology ending in 'omics'. Generally, though not exclusively, these are concentrations of biological molecules such as proteins (proteomics), lipids (lipidomics) or metabolites (metabolomics).

A variety of technological and computational advances in recent years have led to large datasets of molecular omic data becoming available, and these are increasingly being used to investigate complex physiological changes or disease states. The combination of genomics and other omic phenotypes has proven effective for distinguishing between causal and non-causal biomarkers of disease. For example, in the field of proteomics, the OLINK-IMPROVE study (www.olink-improve.com) identified genetic variants regulating plasma protein biomarkers selected based on previous associations with CVD, and identified multiple new causal associations[115]. Recently, an analysis in the Framingham Heart Study, replicated in the INTERVAL and KORA cohorts, performed a similar analysis using 71 proteins previously associated with CVD, finding 69 novel loci, and using Mendelian Randomisation to

identify six proteins which were causal for CHD, two of which were associated with new-onset CHD or CVD events[116]. Proteomic measurements before and after weight loss have been used to identify differential genetic regulation of BMI-associated proteins, which led to the discovery of a new regulator of the satiation hormone leptin[117]. A large-scale lipidomic GWAS identified lipid-associated SNPs which conferred increased risk of coronary artery disease and type-2 diabetes[118], while a similar study using metabolomic measurements found variants linked to a host of diseases including Crohn's disease and CKD[119].

In the case of uric acid, progress has been made in the field of metabolomics by Albrecht *et al.*, who investigated 355 metabolites on the Metabolon platform (www.metabolon.com) by constructing a Gaussian graphical model of significant partial correlations with uric acid in the 1,764 individuals from the KORA cohort[120]. They identified a network of 38 metabolites linked to serum urate level, including nucleotides such as xanthine, a known precursor to uric acid in the purine catabolism pathway, and aspartame, used as a low-calorie sweetener, essential amino acids histidine and methionine and a cluster of steroids along with several uncharacterised metabolites.

Associations between protein concentrations and serum urate have been investigated, but not yet in a high-dimensional analysis – usually, only a few proteins are assessed in any given study. Often, they have only been detected *in vivo*, with physiologically unrealistic concentrations of uric acid, or in species other than humans. While studies in other species are never an ideal way to make inferences about human physiology, they are particularly dubious in the case of uric acid. Uricase inhibitors can be used to reduce the oxidation of urate in model systems, but this can only be an approximation, as human metabolisms have had millions of years to evolve in the presence of high and highly variable levels of serum urate (as loss of uricase activity predates the genus *Homo*), and consequently may respond very differently. Lipids are even more sparsely investigated, with most studies limited to HDL, LDL and total cholesterol and triglyceride levels.

In two of our locally-held cohorts, CROATIA-Vis and ORCADES, we are fortunate to have multiple omic measures available on almost two thousand individuals, including lipids measured by untargeted mass spectrometry and protein concentrations using Olink targeted biomarker platforms (www.olink.com) for inflammation and CVD. Many of these molecular phenotypes have never been investigated in connection with uric

acid before. In the case of the Olink proteins particularly – as serum urate is closely linked to both inflammation and CVD – this may help deconvolve relationships between serum urate levels and the more complex disease phenotypes.

This chapter aims to uncover and interpret novel associations between serum urate and the endophenotypes available in our cohorts. I hypothesise that while many of these phenotypes will be associated with serum urate levels, most of these will be mediated through a much smaller set of associations that persist after accounting for a large number of other phenotypes. Identifying these will highlight candidates for mechanistic associations with serum urate, which may be targets for further research into urate-associated disease.

In addition to simple correlations, multi-phenotype methods will take advantage of the wealth of phenotypic data available. Partial correlations will be used to measure the strength of association between two variables after controlling for the effects of one or more covariates. This reduces the problem of confounding, which is likely to be high in a dataset of biomarkers of related function, and also allows known effects, such as the link between serum urate and BMI, to be taken into account.

As an alternative method to account for high collinearity, lasso regressions will be used to identify minimal sets of urate-predicting phenotypes. Lasso regression is a regression method which penalises the number of non-zero coefficients in the model – effectively, it attempts to identify the simplest possible combination of phenotypes to predict the level of serum urate. By iterating the algorithm multiple times, it is possible to identify a small set of phenotypes whose combination of values best predict serum urate concentration. While partial correlations account for the effect of every phenotype, lasso regression instead discards uninformative phenotypes.

Both methods should identify associations that are independent of other phenotypes in the dataset. Including such a wide range of other phenotypes should increase the chance that those detected are most directly relevant to serum urate levels. It should be noted that these statistical methods cannot account for complex non-linear biological relationships, such as feedback mechanisms, and so should be interpreted cautiously. However, they have the potential to identify new hypotheses which, following experimental verification, may lead to mechanistic insights into either serum urate regulation, or into the impact of serum urate on complex traits, including diseases.

2.1.1 Cohorts

2.1.1.1 CROATIA-Vis

The islands of the Dalmatian coast are home to a number of isolate populations, many of which have been studied for a number of years due to the potential of enrichment of rare genetic variation due to genetic drift and population bottlenecks[121]. The CROATIA-Vis cohort (also referred to as Vis in this thesis) is one of these research populations. It comprises 1,008 participants recruited from the Croatian island of Vis between 2003 and 2004[122], with blood DNA, plasma and serum collected from fasting participants, as well as information from questionnaires (including lifestyle information such as diet) and anthropometric and physical measurements. Collected EDTA plasma was stored at -80°C until the date of Olink analysis in 2016.

2.1.1.2 ORCADES

The Orkney Complex Disease Study, or ORCADES, is another isolate population cohort from the Orkney islands off the north coast of Scotland. Individuals were eligible for recruitment if at least two grandparents resided in the Orkney Isles. A total of 2,080 participants were recruited between 2005 and 2011, with a variety of phenotypes measured by clinic visit and questionnaire. Fasting blood samples were taken from all participants and a subset stored at -80°C until used for Olink proteomics in 2016.

2.1.1.3 INTERVAL

The INTERVAL study was originally set up by the Universities of Cambridge and Oxford to identify the optimum interval for blood donation[123]. 45,000 individuals were recruited between mid-2012 and mid-2014, and Olink proteomic measurement was performed on a subset of 5,000 using blood drawn between 2014 and 2016. Blood samples were generally processed within one day of bleed and frozen. Demographic characteristics were obtained by online questionnaire. Because the sample is comprised of blood donors, participants were in good health at the time of recruitment, but samples were non-fasting.

INTERVAL differs from the other cohorts in the analysis in that serum urate and serum creatinine were measured on the Metabolon platform rather than with standard biochemical assays. This method is a high-throughput mass spectrometry-based system that reports quantities in arbitrary units rather than absolute values. Because the following methods make use of data that has been rank-transformed to normality, this should not affect the results. This assumption has been tested (as detailed in

2.2.5.1) but there may still be differences, and as such caution is warranted when results from INTERVAL are inconsistent with the other cohorts. Because estimated glomerular filtration rate (eGFR) is calculated based on absolute units of creatinine, it is not possible to exclude CKD cases based on eGFR values in this cohort. However, since the cohort was comprised of blood donors, it is unlikely that many, if any, individuals would have been CKD patients.

2.1.1.4 The Estonian Biobank (EGCUT)

The Estonian Biobank is a cohort run by the Estonian Genome Center of the University of Tartu (EGCUT). The EGCUT cohort comprises over 50,000 individuals from Estonia, recruited from the general population by GPs, who performed health examinations as well as taking blood, plasma and DNA[124]. Lifestyle factors were assessed by questionnaire. Olink data is available on a subset of the population. Blood was drawn from non-fasting individuals between 2011-2012 and stored as aliquots in MAPI straws in liquid nitrogen until analysis in May 2017.

2.1.1.5 Lifelines DEEP

Lifelines DEEP is a prospective, general population cohort study from the Netherlands[125]. It is a deeply-phenotyped sub-cohort of the Lifelines cohort comprising 1,539 individuals aged 18 and over. The cohort has extensive molecular phenotyping, which includes Olink proteomics on the CVD-III panel only (see Section 2.1.2.1.1). Fasting blood samples were collected between April and August 2013 and stored at -80°C until the date of analysis in 2018. Uric acid and creatinine are measured by liquid chromatography-mass spectrometry, and so units are not comparable to other cohorts.

2.1.1.6 PIVUS

The Prospective Investigation of Vasculature in Uppsala Seniors (PIVUS) is unique amongst the studies included in this thesis in that all 1,000 participants are the same age – 70 years old at the time of recruitment. The cohort was recruited between 2001 and 2004, with the aim of investigating cardiac function in seniors from the Swedish city of Uppsala. The cohort was recalled for additional phenotyping at age 75 and is currently being updated with age 80 measurements. Olink protein measures are available at age 70 on the CVD-I panel (see Section 2.1.2.1.1), which has some overlap with the phenotypes investigated on our panels. Blood samples were collected after an overnight fast at baseline in 2001-2004 and were frozen without thawing at -80°C until the date of analysis in 2014.

2.1.2 Phenotypes

2.1.2.1 Non-omic phenotypes

Initial investigation included a wide range of phenotypes with possible links to serum urate: body mass index (BMI), A Body Shape Index (ABSI, a measure of body shape designed to assess the effect of body shape independently of BMI[126]), waist-to-hip ratio, systolic and diastolic blood pressure, HDL and LDL cholesterol, triglycerides, total cholesterol, C-reactive protein (CRP), fasting glucose, alcohol consumption (grams/week), insulin, creatinine and eGFR.

Final lipidomic analyses retained all covariates. Olink analyses were limited to BMI, eGFR and alcohol consumption, as these were the only phenotypes available in all replication cohorts. These are categorised as 'non-Olink phenotypes' in all figures and tables, as detailed in Supplementary Table 1.

2.1.2.1.1 Olink proteomics

Olink Proteomics' Olink platform (initially sold under the 'Olink Proseek' name) is a targeted system for quantifying protein biomarkers. This method uses their 'Proximity Extension Assay' technology[127], in which each target protein is bound by pairs of antibodies. Each antibody has a unique DNA oligomer bound that can hybridise only with the second antibody for the target protein. Hybridised sequences are extended and amplified, and quantitative PCR is used to measure the concentration of the hybridised oligomers as a proxy for the concentration of the target protein. Mismatched oligomers do not hybridise and are not detected, improving the specificity of the technology. Ninety-two proteins are measured simultaneously by each panel.

Proteins were measured on the 'Cardiovascular II' (CVD-II), 'Cardiovascular III' (CVD-III) and 'Inflammation' (INF) panels in the Vis, ORCADES, EGCUT and INTERVAL cohorts. A total of 266 proteins were used in the final analyses. Lifelines DEEP used only the CVD-III panel. PIVUS has measurements from the retired 'Cardiovascular I' (CVD-I) panel, which includes a mix of proteins measured on the INF, CVD-II and CVD-III panels. A full list of proteins, abbreviations and names used in analyses is given in **Supplementary Table 1**.

2.1.2.1.2 Lipidomics

Lipids were measured in Vis and ORCADES by electrospray ionisation tandem mass spectrometry (ESI-MSMS) as described in Liebisch *et al.* 1999[128] and 2004[129]. Lipids were identified by class, chain length and number of double bonds.

2.2 Methods

2.2.1 Data processing

2.2.1.1 Non-omic phenotypes

eGFR was calculated from standardised serum creatinine measured in mg/dL using the 4-variable MDRD study equation[130]:

$$\text{eGFR} = 186 \times \text{creatinine}^{-1.154} \times \text{age}^{-0.203} \times 0.742 \text{ (if female)} \times 1.212 \text{ (if AA)}$$

Where AA refers to individuals of African American ancestry (my analyses are exclusively European ancestry, but the term is included here for completeness). Individuals with eGFR < 60 mL/min/1.73m^2 were classified as Chronic Kidney Disease (CKD) cases, as per the classification used in the CKDGen Round IV meta-analyses (see Section 4.2.1.2).

It should be noted that the MDRD equation is no longer the preferred method for calculating eGFR, having been superseded by the CKD-EPI equation[131]. I was unaware of this at the time of analysis, and it is a potential limitation of this work. However, as my own analysis only uses rank-transformed eGFR, I do not believe that using it would greatly change the results beyond changing the classification of a few CKD cases.

BMI was calculated using height in metres and weight in kilogrammes using the standard formula:

$$\text{BMI} = \frac{\text{weight}}{\text{height}^2}$$

2.2.1.2 Olink phenotypes

Protein measurements were provided by Olink Bioscience as Normalised Protein Expression (NPX) values, an arbitrary unit on a $\log_2$ scale. NPX values are calculated using the inverse of the C_t values from the qPCR stage of the Olink process, meaning a high NPX corresponds to a high protein concentration in the sample. Where a measurement was below the Lower Limit of Detection value (LLOD) provided by Olink, that measurement was set to the LLOD value for that protein. **Table 1** lists all proteins with more than 5% of their measurements set to LLOD, based on the original full set of 1,920 participants in the discovery cohort before filtering for complete phenotype data. No proteins were filtered on LLOD percentage, as this may have

differed between replication cohorts and may have resulted in incomplete overlap of phenotypes.

The measurements for BDNF and CCL22 were removed from the dataset as the assays were withdrawn by Olink due to quality issues.

A batch effect was identified in the Olink measurements – linear regression identified a significant effect of sample plate ID on many Olink proteins. To correct for this, plate ID was included as a covariate when adjusting phenotypes (see Section 2.2.1.4).

Table 1 - Olink proteins with more than 5% of samples set to LLOD before filtering.

Protein	% set to LLOD
CVD3_189_KLK6	6.5%
INF_153_Beta.NGF	7.0%
CVD2_190_CA5A	8.3%
CVD2_148_SERPINA12	11.7%
INF_105_MCP.3	14.2%
CVD3_150_PSP.D	14.5%
INF_114_IL.17C	17.0%
INF_141_FGF.5	20.2%
INF_178_IFN.gamma	29.6%
INF_116_IL.17A	34.9%
CVD3_139_SPON1	35.3%
CVD3_195_NT.pro.BNP	58.3%
INF_140_IL.10RA	62.4%
INF_152_PD.L1	67.3%
CVD2_181_BNP	70.7%
INF_193_IL.5	74.4%
INF_181_LIF	79.8%
INF_121_IL.20RA	80.0%
INF_171_IL.20	81.2%
INF_124_IL.2RB	84.4%
INF_163_TNF	84.4%
INF_182_NRTN	85.8%
INF_180_IL.4	86.9%
INF_160_ARTN	88.1%
INF_158_IL.24	89.0%
INF_159_IL.13	90.3%
INF_129_TSLP	90.9%
INF_150_IL.22.RA1	94.6%
INF_127_IL.2	97.2%
INF_125_IL.1.alpha	97.7%
INF_177_IL.33	98.3%

2.2.1.3 Lipidomic phenotypes

Measurements for Vis showed clear signs of batch effects (examples are shown in **Figure 6**), but no information was available on technical covariates to correct for this.

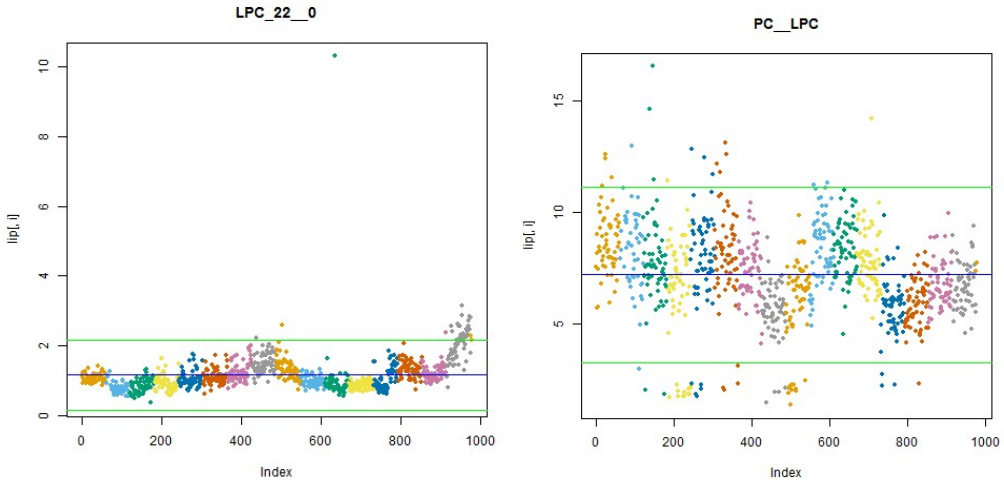


Figure 6 - An example of the technical bias visible in the lipidomics measurements. Colour corresponds to inferred batch number with a batch size of 61.

As an interim solution, it was postulated the technical bias was connected to sample batch, possibly corresponding to analysis day. An algorithm was devised to establish how many samples to assign to each batch. This was based on the theory that each batch would have a distribution of data points around a mean, and that the 'correct' batch would minimise the standard deviation, as it would contain only points drawn from a single distribution.

1. Assign individuals to batches of equal size batch based on ID order
2. Adjust all phenotypes to range between 0 and 1
3. For each phenotype, calculate the standard deviation of measurements within that batch.
4. Take the mean of the standard deviations across all batches.
5. Iterate this across batch sizes from 50 – 100
6. For each batch size, sum the mean standard deviations across all phenotypes.
7. This assigns each batch size a score that represents how well it minimises standard errors across all phenotypes. The batch size with the lowest score is the best performing.

8. For each phenotype, rank the batch sizes based on their mean batch-wise standard deviation. Batch sizes which accurately capture the batch effect seen in the data have low ranks.
9. The best batch size is the one with the lowest mean rank across all phenotypes.

Figure 7A shows the mean batch-wise SD across all phenotypes and batch sizes. The optimum batch size was 61, which can be seen more easily in **Figure 7B**, where it clearly has the lowest mean rank of any batch size. Consequently, individuals were assigned to batches of 61, and this was included in covariate adjustment, as described below.

Figure 7B also compares the batch effect in Vis to that in ORCADES. In Vis, several batch sizes minimise the mean per-batch standard deviation across many phenotypes: 61, 65 and 75. In ORCADES, this is not seen – most batches perform roughly equally, showing that the cross-phenotype batch effect seen in Vis is not present in this cohort.

Unfortunately, when uric acid was regressed on batch number, a significant association was identified for several lipid species. A similar effect was seen when regressing age on batch number, and it was noted that the distribution between sexes was highly uneven. Taken together, these likely explain the effect on serum urate levels.

Given this association between supposed technical batch and phenotypes which were not connected to the lipidomic measurements, I decided to drop the attempt to correct the data, to avoid removing genuine biology, or worse, introducing bias. Lacking a reliable statistical method to separate unreliable-looking phenotypes from reliable ones – judgement by eye being too subjective – I have decided to retain all phenotypes in the models. It is my belief that the inclusion of the cleaner ORCADES data in the analysis will limit the ability of these distortions to drive false positives.

This is an imperfect solution to the problem of bias in the data, but given the lack of accurate technical covariates, I consider this to be the best available option.

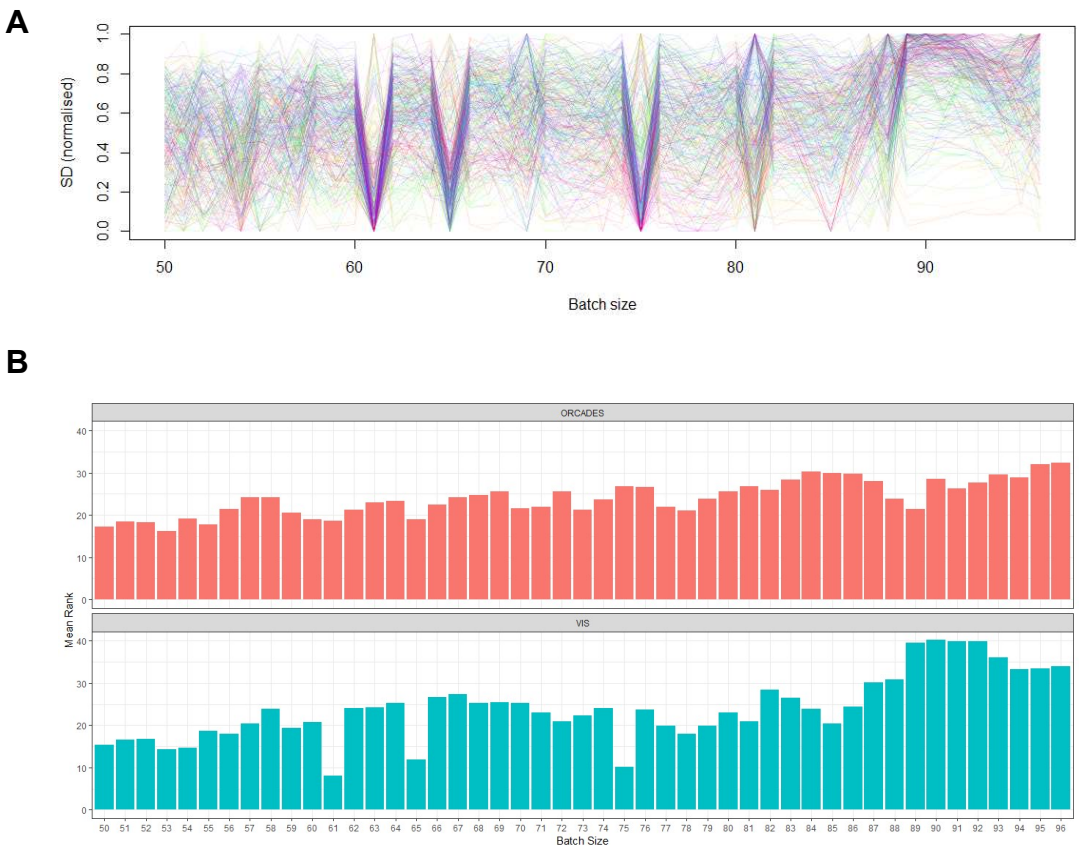


Figure 7 – Comparison of mean standard errors across batch sizes.

A. Mean standard error across all batch sizes.

Each line is a single lipid phenotype in CROATIA-Vis. X-axis is batch size, y-axis is the mean standard deviation across all batches for a given batch size. The highly visible troughs in the graph are batch sizes that minimise the mean standard error for many phenotypes.

B. Mean rank across all phenotypes of each batch size

Lower mean ranks mean the batch size minimises mean SD across more phenotypes. The top panel shows the results for ORCADES, the lower panel for Vis.

2.2.1.4 Adjustment for covariates

Data from the CROATIA-Vis and ORCADES cohorts were merged into a single dataset henceforth referred to as the discovery dataset.

All raw measurements were rank transformed to normality using the 'rntransform' function from the R package *GenABEL*[132]. Linear models were run using the 'lm' function in R (package *stats*, included in base R), regressing normalised phenotypes on age, sex and cohort to remove the effect of these covariates. Olink phenotypes

were additionally regressed on sample plate ID to remove batch effects from the original assay.

The residuals from these models were used as input phenotypes for the correlation, partial correlation and lasso regression analyses; all further reference to 'phenotype' in this chapter refers to these residuals unless otherwise specified.

2.2.1.5 Adjustment for kinship

The discovery cohort was initially adjusted for relatedness by fitting a random genetic effect in addition to the covariates above, using the 'polygenic' function in the *GenABEL* package instead of 'lm'. However, this was not performed in the replication cohorts, as these were composed of unrelated individuals, and while the adjustment led to negligible differences in the partial correlation results, it completely removed the known association with eGFR in lasso regression, which was identified in all other cohorts. Because of this, I decided to drop the kinship adjustment, to ensure consistency between discovery and replication, and to retain this known biologically relevant association in the models.

2.2.2 Correlations

Spearman's rank correlations were calculated between serum urate and all phenotypes using the 'rcorr' function from the *Hmisc* R package[133], which additionally returns p-values. Bonferroni correction was used to correct for multiple testing (269 tests for Olink), resulting in a significance threshold of $p < 1.86 \times 10^{-4}$. Correlations were calculated for the whole dataset, for both sexes separately and sex-combined with CKD cases excluded. Differences in effect between sexes were tested using Fisher's r-to-z transformation.

2.2.3 Partial correlations

Partial correlations were calculated using the *ppcor* package[134] for R, which also returns p-values for each partial correlation. Partial correlations were calculated using the Spearman's rank method between serum urate and all phenotypes in the analysis. False Discovery Rate (FDR) correction (implemented in the 'p.adjust' function in R) was used to adjust p-values for multiple testing (269 tests for Olink proteomics data, 191 for lipidomics) A significance threshold of $Q < 0.05$ was used to identify significant partial correlations.

2.2.3.1 Partial correlation network

To provide context to the correlated phenotypes, partial correlations were also calculated for each phenotype that was significantly associated with uric acid. These primary and secondary correlations were then converted to a network, where each vertex is a phenotype and each edge represents a significant partial correlation between two phenotypes. Any phenotype with a partial correlation p-value < 0.05 (significant before multiple testing correction) with serum urate was highlighted in the network, as these are most likely to provide relevant context to the relationship.

Networks were created using the *igraph* R package[135] and plotted using *Cytoscape v3.5.1*[136]. All sets of phenotypes associated with a given urate-associated phenotype were also tested for gene ontology term enrichment against the set of all Olink proteins using the GOrilla[137,138] web tool.

2.2.3.2 Sensitivity analysis

Before replication data was available, a sensitivity analysis was performed to provide a measure of confidence that the partial correlations detected were not purely artefacts of the data. Partial correlations were recalculated 1,000 times, randomly removing 20% of the dataset each time, and each phenotype scored for the number of times it was identified as a significant partial correlation with serum urate.

To further test the robustness of the identified correlations, random sampling with replacement was used to create a bootstrapped dataset of equal size to the full discovery. This was then used to calculate partial correlations. This was again repeated 1,000 times, with phenotypes scored for the number of times they were included as a significant partial correlation.

2.2.4 Lasso regression

As an alternative method for identifying phenotypes with a significant association with uric acid levels, lasso regression models were run using uric acid as the response variable and other phenotypes as predictors. Each regression model assigns all phenotypes a coefficient, the majority of which will be zero. Regressions were run using the *glmnet* R package[139]. Regressions were run 1,000 times, to account for the stochasticity inherent in the algorithm, and the mean non-zero coefficient per phenotype recorded, along with the number of times it was included in the final model.

2.2.5 Replication

I identified potential collaborators for replication through the SCALLOP consortium, a group set up to conduct meta-analyses of GWAS for Olink proteomic measurements. Two studies were identified that had serum urate measurements as well as the CVD-II, CVD-III and INF panels – the EGCUT cohort in Estonia and the INTERVAL study in Cambridge. Non-omic phenotypes included in the analysis were reduced to match those available in the replication cohorts as described in section 2.2.1.1. Additionally, the Lifelines-DEEP cohort analysed correlations with the CVD-III panel only. Sex-separate partial correlations and lasso regression were not analysed in EGCUT due to insufficient sample size. Additionally, exclusion of CKD cases from INTERVAL was not possible, due to the arbitrary units of creatinine measurement making the calculated eGFR values incomparable to the normal classification scheme.

I provided replication analysts with an R script to align and automate the processing and analysis of their data.

2.2.5.1 Checking INTERVAL results

In the INTERVAL cohort, serum urate and serum creatinine were measured on the metabolon platform, rather than using traditional assays. Because of this, I performed additional tests to confirm the results in this cohort.

Metabolon data is also available in the ORCADES cohort, allowing comparison with the biochemistry measures. Serum urate measured with metabolon had a Spearman correlation of 0.874 with clinical biochemistry laboratory-measured serum urate, while creatinine had a corresponding correlation of 0.793. Considering that the original measurements were made shortly after collection, but the Metabolon analysis was performed 6-12 years later on samples that had been stored at -80°C, I consider this correlation to be high.

Because eGFR is sensitive to creatinine units, and has a non-linear relationship with creatinine values, ‘eGFR’ calculated in from Metabolon creatinine in ORCADES did not correlate with true eGFR. To test the suitability of raw creatinine as a replacement, the above analyses were rerun in both the discovery cohort and INTERVAL using creatinine in place of eGFR, and in both cases the results were identical to four decimal places. This means that although there is not a 1:1 mapping between eGFR and metabolon-eGFR, the results of the analysis do not appear to be affected by this.

2.2.6 Serum urate GWAS lookups

A protein quantitative trait locus (pQTL) is a region of the genome containing variants which are associated with levels of a protein. If a pQTL is also significantly associated with a trait of interest, it can be indicative that the protein in question is linked to that trait. In the case of a cis-pQTL (a pQTL in the region of the gene encoding the protein), it supports a causal role for the protein.

2.2.6.1 Vis and ORCADES meta-analysis

To identify pQTLs for the Olink proteins in our data, I made use of meta-analyses of Olink proteins in Vis and ORCADES (individual GWAS run performed by Andrew Bretherick, University of Edinburgh, meta-analysis by Anne Richmond and Thibaud Boutin, in our group). I defined pQTLs in this data using the genome-wide significance threshold of $5x10^{-8}$, and additionally using the commonly accepted threshold for suggestive significance ($P < 1x10^{-5}$) as the power of the GWAS was limited by sample size.

For this analysis, I defined a locus as a region of 1Mb centred around the index SNP, the SNP with the lowest suggestively significant P-value in a chromosomal region. Index SNPs were then looked up in the CKDGen serum urate transethnic meta-analysis (see 4.2.2) to check for a significant association with serum urate levels.

SNPs were additionally looked up in the PhenoScanner database[140], which includes data from the NHGRI-EBI GWAS catalogue[141], NHLBI GRASP[142] and NCBI dbGaP[143] catalogues in addition to 137 GWAS datasets generated by the developers. This reports any published associations between a given SNP and all phenotypes in the database – the GWAS, pQTL and eQTL databases were queried, using a significance threshold of $P < 5x10^{-8}$.

2.2.6.2 OLINK-IMPROVE CVD-I GWAS lookup

Under the OLINK-IMPROVE project, Folkersen *et al.* have published summary statistics from meta-analyses of 3,394 individuals for all proteins on the Olink Cardiovascular Disease 1 (CVD-I) platform[115]. The larger sample size in this meta-analysis compared to ours confers greater power to detect associated SNPs, but due to the use of a different Olink panel, only four of the proteins identified as correlated with serum urate have association signals in this data: FGF-23, CCL3, FABP4 and CHI3L1.

Initially, the index SNPs identified for these proteins were also looked up as above, except that the standard GWAS threshold for significance (5×10^{-8}) was used to select SNPs. However, for FABP4 and CCL3 an excessively large number of loci (of the order of 50 and 100 respectively) were identified using the provided summary statistics, while the accompanying publication lists only two pQTLs for CCL3 and none for FABP4. As this is far in excess of the number of significant loci I would expect for an endophenotype GWAS with a sample of this size I have sought clarification from the authors of the study on whether this is an error in the database, or whether additional filtering is required. At time of writing, this has not been resolved.

For the current analysis, the pQTL index SNPs reported in the paper for CCL3 and CHI3L1 have been used for lookup, as well as two SNPs for FGF-23 which were present in the summary statistics, but not explicitly referenced in the paper.

2.2.6.3 SOMAscan lookup

The INTERVAL study has published pQTLs for 3,622 plasma proteins in 3,301 healthy participants, measured on the SomaLogic SOMAscan platform (www.somalogic.com). Rather than using antibodies, the technology uses DNA-aptamers to bind proteins and measure their abundance. This allows it to cover a wider range of proteins than Olink, but it ultimately remains a targeted approach. Full GWAS results have been made available online, including results for FGF-23, CCL3, CHI3L1, FABP4, IGFBP-2, LDL-receptor and MMP2. No results were available for Ep-CAM, FGF-21, PLC or PON3.

Index SNPs were identified using the 5×10^{-8} threshold for genome-wide significance and 1Mb window, as in Section 2.2.6.1, and the SNPs looked up in the serum urate meta-analysis results.

2.2.6.4 Colocalisation analysis

A SNP associated with multiple traits can be an indication of a mechanistic link between those two traits. An apparent shared signal can arise from linkage disequilibrium (LD) between two separate causal signals or through genuine pleiotropy – a causal signal having an effect on two otherwise independent traits. Distinguishing between these two scenarios is useful when interpreting associations, as a shared causal SNP is much stronger evidence for a shared aetiology.

Colocalisation analysis is one way to compare signals from two traits. This is a Bayesian statistical method using two sets of GWAS summary statistics to calculate

the posterior probabilities of five scenarios: H_0, the null hypothesis, is that there is no signal in a region for either trait. H_1 and H_2 are the hypothesis that there is one causal SNP associated with either trait one or trait two only. H_3 is the hypothesis that there are two separate causal SNPs, one associated with each trait. The final hypothesis, H_4, is the scenario representing true colocalisation – one causal SNP affecting both traits.

Colocalisation analyses were performed using the 'coloc.fast' function in the *gtx* R package[144], which implements a modified version of Giambartolomei's colocalisation method[145]. This method does not require individual-level genotypes, taking only GWAS summary statistics as input. It fits a model assuming that of all SNPs in the dataset, a maximum of one can be truly causal for each trait. The method integrates across all possible combinations of causal SNPs and calculates a posterior probability for the data observed under each hypothesis. The prior probability of a SNP being associated with a trait was set at the default for the function: 10^{-4} for each SNP being causal for one trait, and 10^{-5} for being causal for both. A posterior probability of ≥0.8 was taken as the threshold for accepting a hypothesis. The window was defined as the region of 1Mb centred around the index SNP. All SNPs within this window present in both sets of summary statistics were included in the calculation.

It must be caveated that the method assumes the two populations are unrelated and drawn from the same ethnic group – in this case, the Olink meta-analysis populations, Vis and ORCADES, have also contributed to the serum urate meta-analysis. However, as these two studies contribute fewer than 3,000 individuals to a study containing over 450,000 samples, the impact of this overlap should be minimised. Furthermore, the transethnic serum urate GWAS results were used, as I did not have access to the European-ancestry summary statistics at the time the colocalisation analysis was run.

2.2.7 GENOSCORES

The GENOSCORES software package[146] uses GWAS summary statistics and population genotype data to calculate genetic risk scores on a per-locus basis for a trait. Though not the primary aim of the software, it can be used to obtain a measure of a regional genetic correlation, achieved by calculating multiple scores for a single analysis population and obtaining the correlation between them.

For all SNPs in a locus, the genotypes of the analysis population are weighted by the GWAS effects and summed to obtain a score. For each pair of traits, a correlation can then be calculated for all regions on the same chromosome.

This differs from more traditional genetic risk scores in that it focuses on all SNPs in a single region, rather than a single index SNP from all regions across the genome, and from LD-score regression based genetic correlations in that it considers each region separately. In this way, it attempts to capture the modularity of complex traits.

The summary statistics from the Olink meta-analysis described in Section 2.2.6 were uploaded to the GENOSCORES database. These were then used by Athina Spiliopoulou in conjunction with the summary statistics for the Köttgen *et al.* 2012 GWAS of uric acid[74] and the genotypes from to calculate genetic score correlations between serum urate and Olink proteins for the 503 individuals in the 1000 Genomes Project European subset[147]. Scores for Olink were compiled using all SNPs with a P-value of less than 1×10^{-5}, with separate scores for *cis* and *trans*-eQTLs. I used these pre-calculated scores to provide additional context to the serum urate-Olink associations reported in this chapter.

2.2.8 Genetic correlation with LD-score regression

Genetic correlation reflects the amount of covariance between two traits due to genetics. Essentially it is a measure of the extent to which two phenotypes are explained by common genetics – the phenomenon known as pleiotropy. The existence of a genetic correlation between two phenotypes can be evidence of a mechanistic link between them, and so it is a valuable tool in genetic analysis. Genetic correlation is expected to be strong if there is a causal relationship between an intermediate trait and an outcome of interest.

It has an advantage over Mendelian randomisation (MR) in that it does not use only significant SNPs – this makes it more useful in cases where a significant fraction of the heritability of the trait cannot be explained by multiple significant SNPs. The disadvantage is that genetic correlation is a necessary but not sufficient condition for a causal relationship. A genetic correlation between two traits is not definitive proof of causality – any apparent correlation can be driven by an unmeasured trait that is the true mediator of the genetic effects on one or both of the measured traits. However, the problem of confounding is considerably less than for straight phenotype-

phenotype correlations, as any unmeasured trait must be genetic rather than environmental in origin.

Cross trait LD-score regression is a method developed by Bulik-Sullivan *et al.*[148] that uses summary-level GWAS results to estimate genetic correlation between two traits. It can be applied to summary-level data if a reference population of similar ancestry is available to determine LD pattern, meaning access to individual level genotypes is not required.

The method is an extension of single-trait LD-score regression[149], where GWAS χ^2 statistics are regressed against LD-score, a measure of the extent to which a SNP is in LD with its neighbours, to estimate SNP heritability. The method can be extended by substituting the product of the Z scores for two traits for each SNP in place for χ^2. In this scenario, the regression coefficient is proportional to the genetic correlation coefficient r_g.

Cross-trait LD-score regression was performed using both the *LDSC* software package[148,149] and the LD-Hub website[150], an online database and analysis tool that aggregates published GWAS summary statistics for the purpose of calculating genetic correlations.

2.3 Olink Results

2.3.1 Cohort summary

Table 2 summarises cohort size, sex and CKD status. **Table 3** contains phenotype summary statistics broken down by cohort, prior to any adjustment or transformation.

Table 2 – Sample sizes and phenotype summary statistics for serum urate-Olink correlation analyses.

Cohort	n	% male	% CKD
CROATIA-Vis	698	41.69%	8.02%
ORCADES	881	45.29%	2.38%
EGCUT	475	52.21%	3.16%
INTERVAL	687	59.53%	-
Lifelines DEEP	1,052	42.30%	0.48%
PIVUS	827	50.42%	3.51%

Table 3 – Non-Olink phenotype summary statistics.

Phenotype	Cohort	Mean	SD	Median	Min.	Max.	Notes
Age	Vis	56.62	15.22	57.00	19.00	91.00	
	ORCADES	52.39	15.12	52.65	17.12	91.47	
	EGCUT	53.87	14.10	55.00	23.00	87.00	
	INTERVAL	59.54	6.26	58.80	48.90	76.20	
	Lifelines DEEP	45.23	13.47	45.54	18.00	81.42	
	PIVUS	70.15	0.15	70.14	69.80	70.71	All cohort participants age 70 or very close
Alcohol Consumption (g/week)	Vis	99.18	164.74	30.56	0.00	1500.00	
	ORCADES	85.51	109.14	42.00	0.00	1176.00	
	EGCUT	47.65	117.01	12.55	0.00	1159.20	
	INTERVAL	71.34	58.24	48.00	1.60	208.00	
	Lifelines DEEP	76.77	117.92	45.01	0.00	1300.60	
	PIVUS	46.83	53.87	30.30	0.00	428.80	
BMI (kg/m^2)	Vis	27.41	4.21	27.31	17.08	43.60	
	ORCADES	27.64	4.88	26.91	16.97	51.11	
	EGCUT	28.54	5.66	27.76	17.30	52.77	
	INTERVAL	26.58	4.22	26.06	16.26	43.99	
	Lifelines DEEP	25.18	4.06	24.58	16.67	44.92	
	PIVUS	27.02	4.25	26.57	16.56	49.77	
eGFR	Vis	87.99	22.64	86.82	9.27	214.58	
	ORCADES	100.08	24.00	96.71	26.98	243.57	
	EGCUT	102.91	25.11	101.25	40.29	209.15	
	INTERVAL	148,43.08	12,431.20	13,814.39	5,979.95	327,491.29	Calculated with arbitrary units of creatinine
	Lifelines DEEP	112.05	23.35	109.93	41.32	199.40	
	PIVUS	92.50	19.18	90.75	24.88	188.28	
Serum uric acid (mg/dL)	Vis	5.19	1.59	5.07	1.24	11.57	
	ORCADES	4.97	1.16	4.94	1.18	9.15	
	EGCUT	5.38	1.49	5.23	1.16	10.66	
	INTERVAL	1.07	0.25	1.06	0.46	1.96	Measured on Metabolon in arbitrary units
	Lifelines DEEP	1,819,993.00	369,571.70	1,773,099.00	952,953.00	3,278,282.00	Measured on LC-MS platform
	PIVUS	5.73	1.37	5.59	2.26	10.81	

2.3.2 Correlations

Correlations between all phenotypes are shown in **Figure 8**. Most correlations are positive, and there are clear inter-panel correlations, particularly on CVD-III. Because Olink phenotypes are biomarkers selected for their connection to cardiovascular or inflammatory diseases or biological processes, some correlation is to be expected.

Correcting for the panel effect would not be straightforward, and would risk removing actual biological effects, so these have not been adjusted for.

Forty-nine phenotypes had significant correlations with serum urate in at least one discovery sub-analysis after Bonferroni multiple testing correction ($p < 0.000173 = 0.05/289$) (**Figure 9**). Of these, 43 replicated in at least one replication sub-analysis ($p < 0.00102 = 0.05/49$). Interestingly, CVD-II and CVD-III have many more correlated phenotypes than INF, despite the panels all being of similar size.

In the whole cohort discovery, the strongest correlation identified was with BMI (r 0.314, SE 0.024, $p < 2x10^{-16}$); the strongest Olink correlation was with Leptin (LEP) (r 0.275, SE 0.024, $p < 2x10^{-16}$), an anti-obesity hormone known to correlate positively with BMI in healthy individuals. The strongest negative correlation was with IGFBP-2 (r -0.215, SE 0.025, $p < 2x10^{-16}$). **Figure 9** shows a heatmap of correlations for all phenotypes that were significant in at least one discovery sub-analysis.

Excluding CKD cases identified additional significant correlations with matrix metalloproteinase 2 (MMP-2), which replicated in Lifelines DEEP and IL-18, which did not replicate.

Tests for significant differences in correlation coefficient between sexes identified four Olink proteins after multiple testing correction ($p < 0.000173 = 0.05/289$). These are listed in **Table 4**.

Table 4 – Phenotypes with significantly different correlations with SUA between sexes ($n_{female} = 1083$, $n_{male} = 809$)

Phenotype	r_{female}	r_{male}	P_{diff}
CVD2_113_IL.6	0.254	0.077	9.53E-05
CVD2_120_TRAIL.R2	0.276	0.096	6.40E-05
CVD3_114_CSTB	0.297	0.124	9.84E-05
CVD3_160_PRTN3	0.115	-0.070	6.79E-05

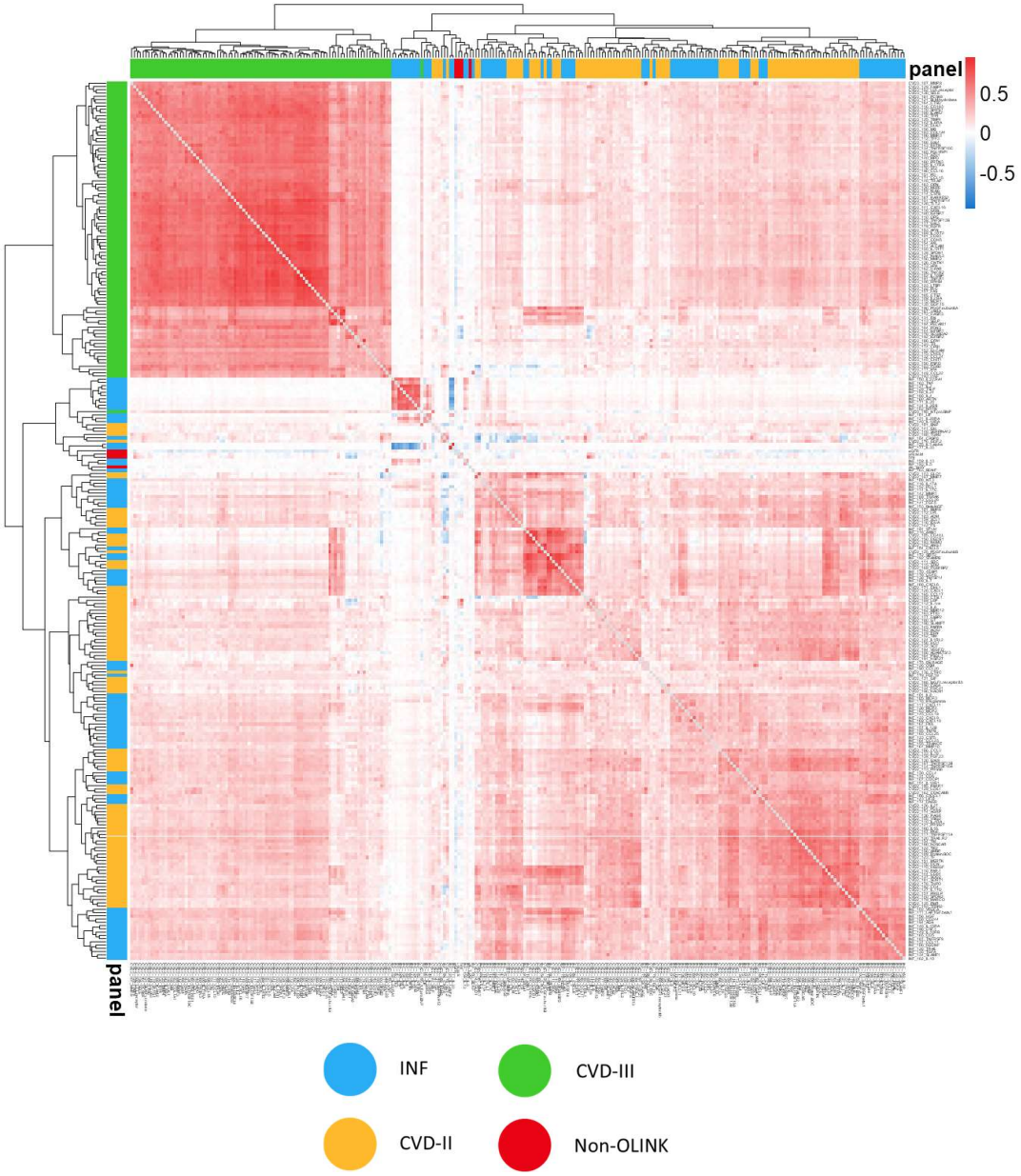


Figure 8 - Heatmap of correlation coefficients between all phenotypes

Spearman correlations were calculated on residuals after adjusting rank-normalised phenotypes for age, sex and Olink plate ID. “Non-Olink” category comprises BMI, eGFR, and alcohol consumption (see Supplementary Table 1 for complete list of phenotypes including Olink panel manifests).

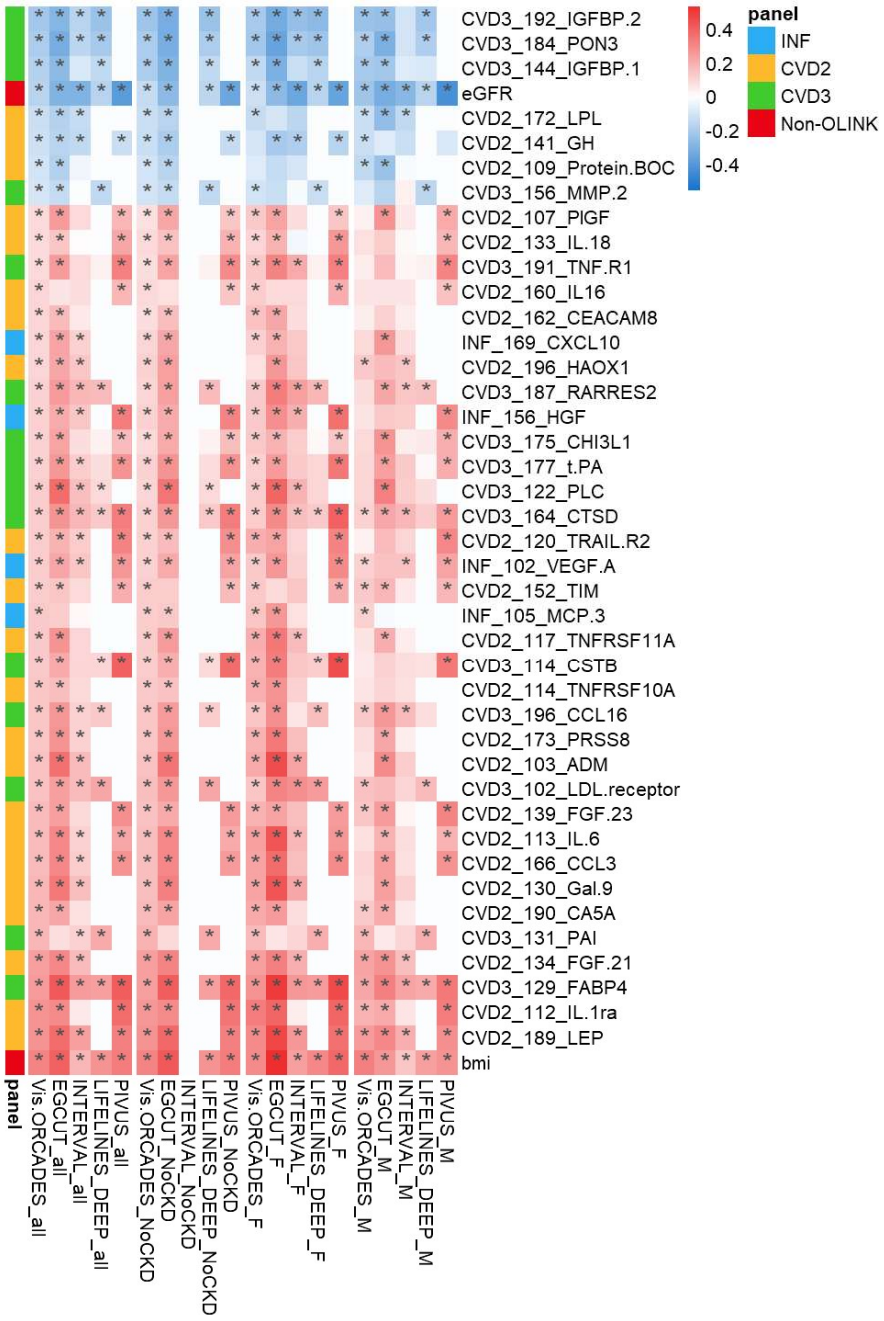


Figure 9 - Heatmap of Spearman's rank correlations between serum urate and Olink phenotypes.

Cell colour corresponds to direction of correlation and intensity to magnitude. '*' denotes a significant P-value. In the case of discovery (Vis.ORCADES), this is P<0.05/289. For replication analyses this is P < 0.05/49. X-axis is grouped by sub-analysis and then by cohort. Y axis is sorted by increasing correlation coefficient in the complete discovery analysis. Only phenotypes with a significant correlation with serum urate in the discovery cohort in at least one analysis that replicates in at least one discovery cohort are shown.

2.3.3 Partial correlations

In the full discovery cohort, significant partial correlations (Q < 0.05) with serum urate were identified with three Olink proteins: Fibroblast Growth Factor 23 (FGF-23), Epithelial Cell Adhesion Molecule (Ep-CAM) and Insulin-Like Growth Factor Binding Protein 2 (IGFBP-2). Excluding CKD cases identified only FGF-23 and Ep-CAM as significant, though the differences in correlation coefficients are small, and the loss of IGFBP-2 may be simply due to smaller sample size.

Directions of correlations were consistent in males and females for all three phenotypes, although none were identified as significant, likely due to small sample size. In FGF-23, partial correlation coefficients were smaller in sex-stratified results, which may indicate a bias due to sex.

Table 5 – Partial correlations in the discovery sample.
Includes all phenotypes which are significant after multiple testing correction in any sub-analysis. **Bold** indicates significance in a sub-analysis.

Phenotype	Sub-analysis	Partial correlation coefficient	Std. Err.	P-value	Q-value
FGF-23	**All**	**0.109**	**0.028**	**7.81×10^{-5}**	**0.01**
	CKD-excluded	**0.114**	**0.029**	**6.35×10^{-5}**	**0.01**
	Female	0.077	0.040	0.0556	0.50
	Male	0.088	0.049	0.0725	0.88
Ep-CAM	**All**	**0.100**	**0.028**	**2.69×10^{-4}**	**0.02**
	CKD-excluded	**0.101**	**0.029**	**3.83×10^{-4}**	**0.03**
	Female	0.111	0.040	5.45×10^{-4}	0.25
	Male	0.085	0.049	0.0827	0.88
IGFBP-2	**All**	**-0.104**	**0.028**	**1.53×10^{-4}**	**0.01**
	CKD-excluded	-0.094	0.029	9.37×10^{-4}	0.06
	Female	-0.105	0.040	8.85×10^{-3}	0.30
	Male	-0.082	0.049	0.0945	0.88

2.3.3.1 Partial correlation network

Partial correlations were calculated for the three phenotypes identified in the discovery analyses, and the results plotted as a correlation network (**Figure 10A**). IGFBP-2 was significantly partially correlated with a much larger number of phenotypes than either FGF-23 or Ep-CAM. Removing CKD cases from the sample

removed IGFBP-2, as above, and additional partial correlations were detected between FGF-23 and Selectin-E (SELE), and Ep-CAM and Galectin-4 (Gal-4) (**Figure 10B**).

Several of the phenotypes in the network had significant partial correlations with serum urate before FDR-correction ($p < 0.05$). These are highlighted in the network and may provide the most relevant context for the three significantly-correlated phenotypes.

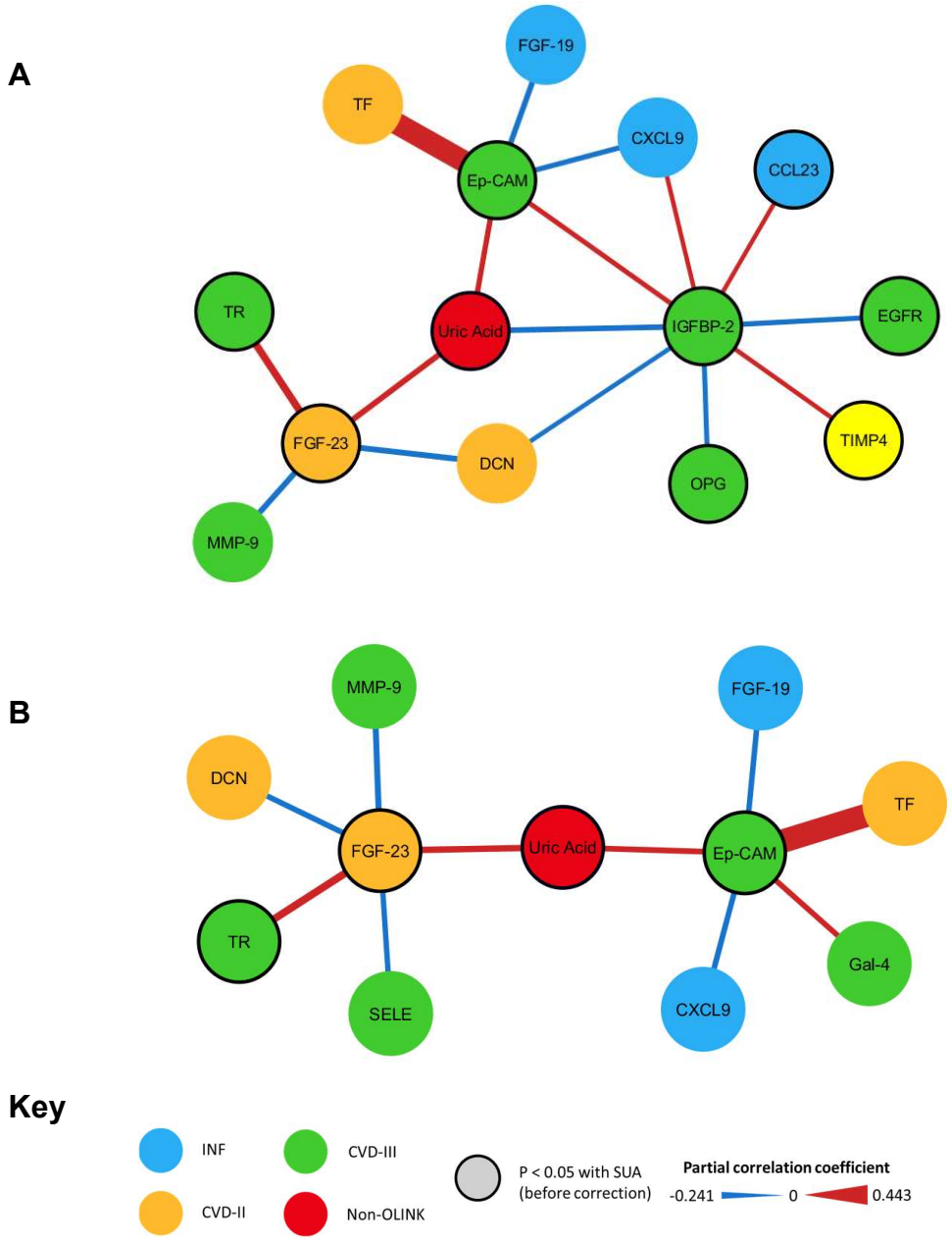


Figure 10 - Partial correlation networks

A: the whole discovery cohort and **B:** with CKD cases excluded. Vertices are phenotypes, colour denotes Olink panel. Vertices with black borders have significant partial correlations with serum urate before multiple testing correction ($P < 0.05$). Edges represent a significant partial correlation between two phenotypes ($Q < 0.05$), (note that partial correlations were only calculated for serum urate, FGF-23, Ep-CAM and IGFBP-2). Edge thickness is proportional to partial correlation coefficient magnitude. Red edges denote positive correlations, blue edges denote negative. (Due to the large number of phenotypes partially correlated with IGFBP-2, nodes are shown only for phenotypes which also had a partial correlation $P < 0.05$ with serum urate).

2.3.3.1.1 Gene Ontology enrichment

No enrichment for any gene ontology category was detected for the set of all proteins included in the network, the set of proteins partially correlated with FGF-23 or the set correlated with EpCAM. The set correlated with IGFBP-2, which is by far the largest set, showed significant enrichment for the 'negative regulation of endopeptidase activity' GO term (Enrichment: 4.34, $P = 8.24 \times 10^{-4}$) but this was not significant after FDR-correction for multiple testing.

2.3.3.2 Sensitivity analysis

Only the three phenotypes detected in the full analysis were identified in more than 10% of the iterations (**Figure 11A**). The only other phenotype to appear in more than 5% of the iterations was ST2 (ST2 Protein).

Many more phenotypes appeared in the analysis of the bootstrapped sample (**Figure 11B**), as would be expected from a sample that is likely to contain duplicated individuals. The most commonly identified phenotypes are FGF-23, Ep-CAM and IGFBP-2, all appearing in nearly 800 iterations, with ST2 appearing in over 500 iterations.

In both analyses there is a clear separation between the number of times FGF-23, Ep-CAM and IGFBP-2 were identified and all other phenotypes, although this gap is narrower in the bootstrapped sample.

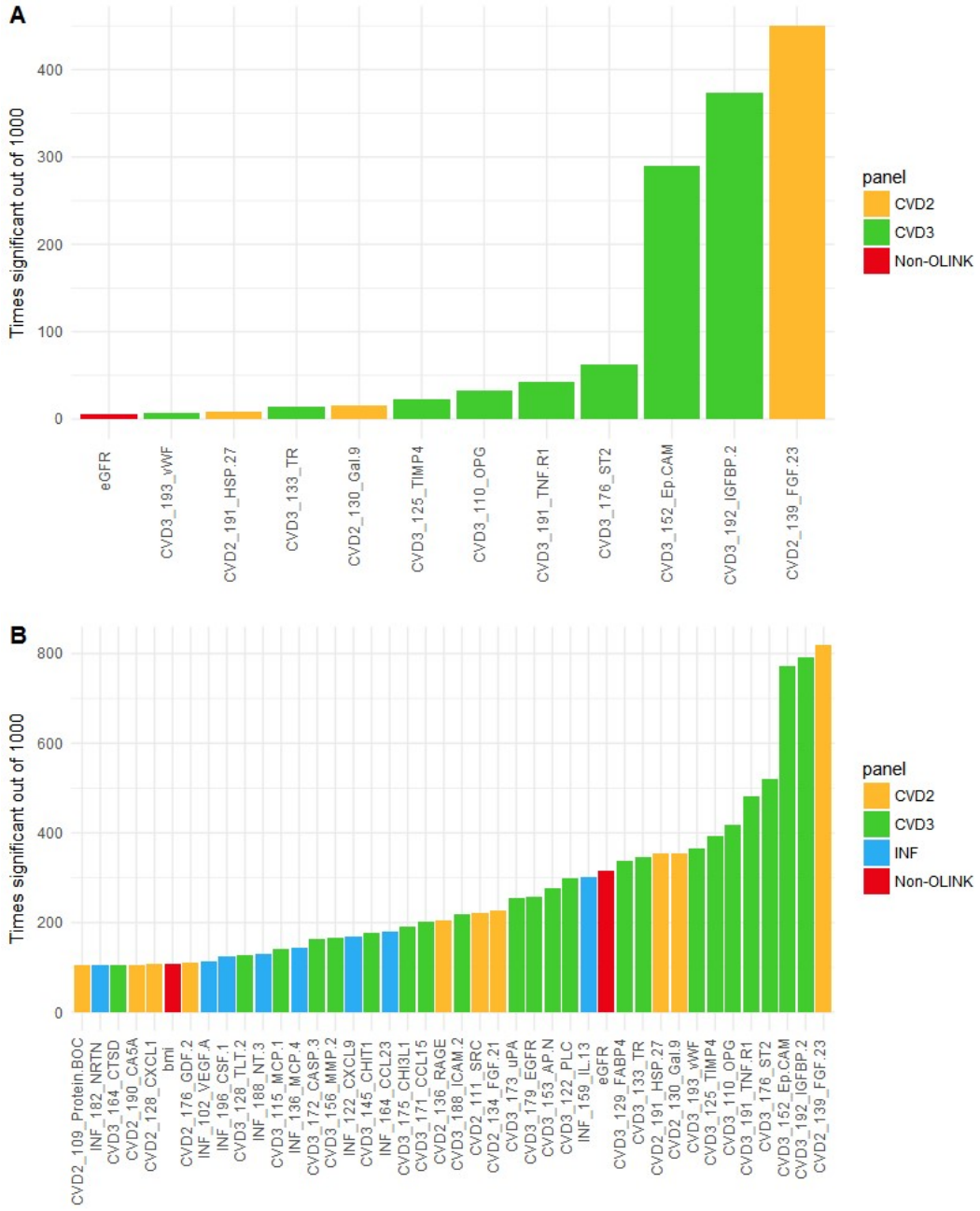


Figure 11 – Partial correlation sensitivity analysis results for whole discovery.
Plots show the number of times a phenotype was identified as a significant partial correlation in 1,000 iterations using either **A**: a random 80% of the full discovery sample (only phenotypes appearing 5 or more times are shown), or **B:** a randomly generated dataset of equal size to the full discovery, generated using random sampling with replacement (only phenotypes appearing 100 or more times are shown).

2.3.3.3 Replication

The results of the partial correlation replication are shown in **Figure 12**. Partial correlation results from Lifelines DEEP and PIVUS are also shown, although only for comparative purposes, as the correlations were only adjusted for proteins on CVD-III and CVD-I respectively.

None of the three significant partial correlations identified in the full discovery were significantly partially correlated with serum urate after multiple testing correction, but this is likely at least partly due to reduced sample size in replication cohorts. However, FGF-23 is significant before multiple testing correction in PIVUS and matches in direction and approximate magnitude in both EGCUT and INTERVAL (as the protein was measured on CVD-II no data from LIFELINES Deep is available). Ep-CAM matches in direction in INTERVAL only. IGFBP-2 has a consistent negative correlation in EGCUT, Lifelines DEEP and PIVUS, although none are significantly different from zero.

Additionally, eGFR was identified as a partial correlation in both INTERVAL and PIVUS. Though serum creatinine and serum urate were measured on Metabolon in INTERVAL, this association seems reliable as it replicates in PIVUS, and is significant before multiple testing correction in both the discovery and the LIFELINES Deep cohorts.

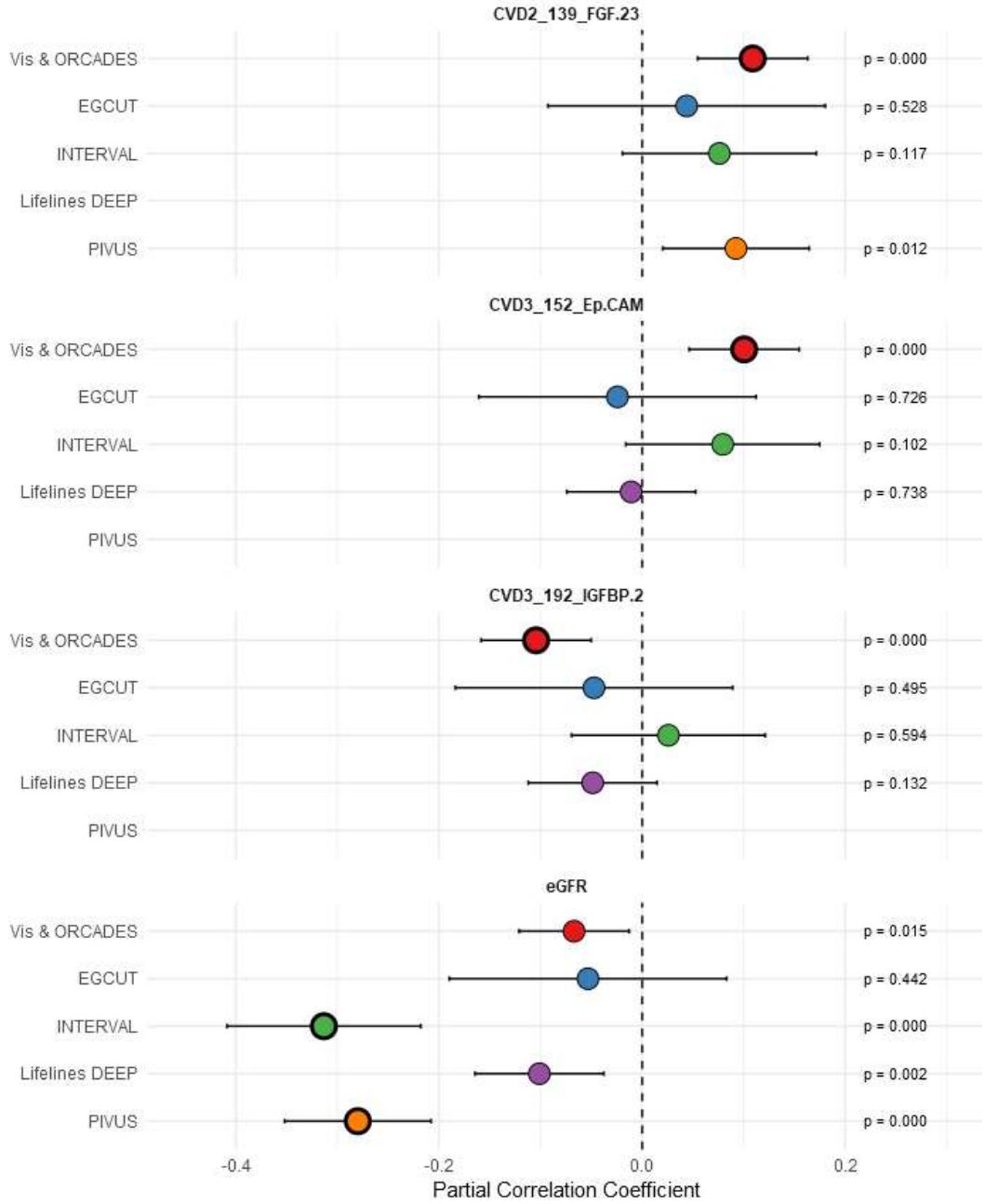


Figure 12 – Partial correlation replication.

Partial correlation coefficients for all phenotypes significant in the full discovery analysis for any cohort. Thick borders indicate significance after FDR-correction for multiple testing in that cohort. Error bars are 95% confidence intervals. Lifelines DEEP only included Olink proteins from the CVD-III panel, and PIVUS only overlapping proteins from the CVD-I panel. As such these should be compared to the other cohorts with caution.

2.3.4 Lasso regression

A phenotype was considered to be identified as consistently predictive of serum urate levels by the lasso regression approach if it had a non-zero coefficient in > 500 iterations in at least two cohorts. The results are shown in **Figure 13**, which compares the mean regression coefficient over all 1000 runs for all urate-predictive phenotypes. In all cases, the sign of the coefficient is consistent between cohorts (where it is non-zero). Full numerical data is shown in

Supplementary Table 2.

Of the phenotypes identified in the partial correlation analysis, FGF-23, IGFBP-2 and eGFR are identified as urate-predictive, with IGFBP-2 assigned the largest absolute coefficient of any phenotype (-0.137). eGFR is included in all models for all cohorts, while IGFBP-2 is present in all models for the discovery and LIFELINES Deep, and 85% of models with a smaller mean coefficient in EGCUT. FGF-23 is only included in models for the discovery and PIVUS cohorts, but in both cases, it is retained in all 1000 iterations. For all three phenotypes, the sign of the coefficients matching the sign of the partial correlation coefficient.

The only phenotype other than eGFR identified in all cohorts that it was included in is Fatty Acid Binding Protein 4 (FABP4), which was positively associated with serum urate (0.101 in discovery). BMI and PON3 were identified in all models for three cohorts, and IGFBP-2 and FGF-21 in two with a third identifying them a large percentage of the models.

Figure 14 shows all phenotypes included in more than 95% of regression models in any cohort. As observed in the correlation analysis (Section 2.3.2) CVD-II and CVD-III again have many more urate-predictive phenotypes than INF.

48 phenotypes were identified in the discovery cohort, 10 in EGCUT and PIVUS, 7 in Lifelines DEEP and only 3 in INTERVAL. This may partly be a consequence of sample size, and, in the case of PIVUS and Lifelines DEEP, the reduced set of phenotypes included in the regression. However, in INTERVAL, serum urate and creatinine were measured on the same Metabolon platform, which may mean that uncorrected technical correlations are driving the strong association between serum urate and eGFR and masking the effect of any Olink associations.

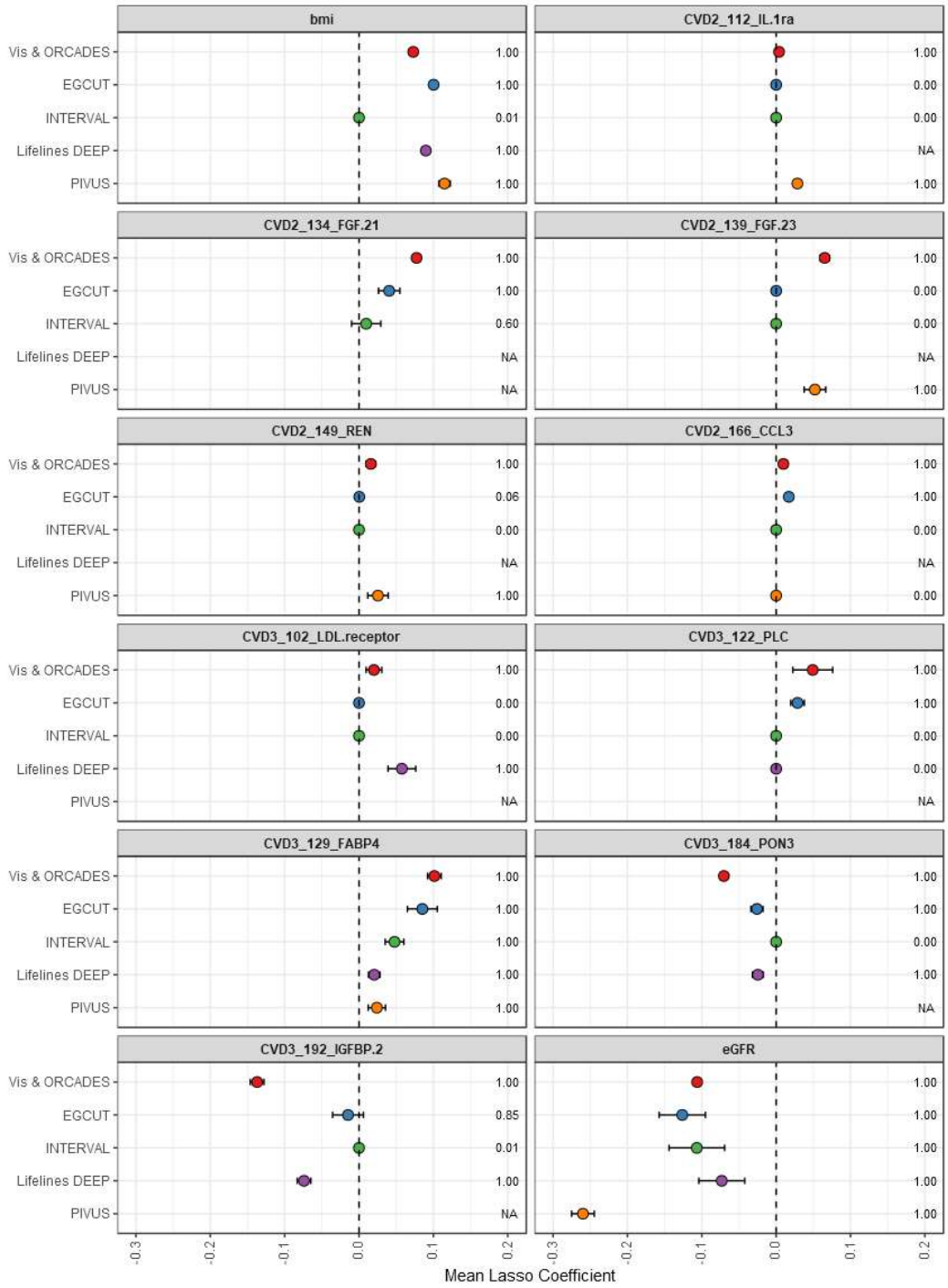


Figure 13 - Mean lasso regression coefficients.

Plotted for all phenotypes retained in the regression model in > 950 iterations in at least two cohorts. Error bars are ± 2 standard deviations (in some cases this is a very narrow range and the error bars are masked by the point). The numbers on the right hand side of each plot are the proportion of times out of 1,000 that the phenotype had a non-zero coefficient in the regression model.

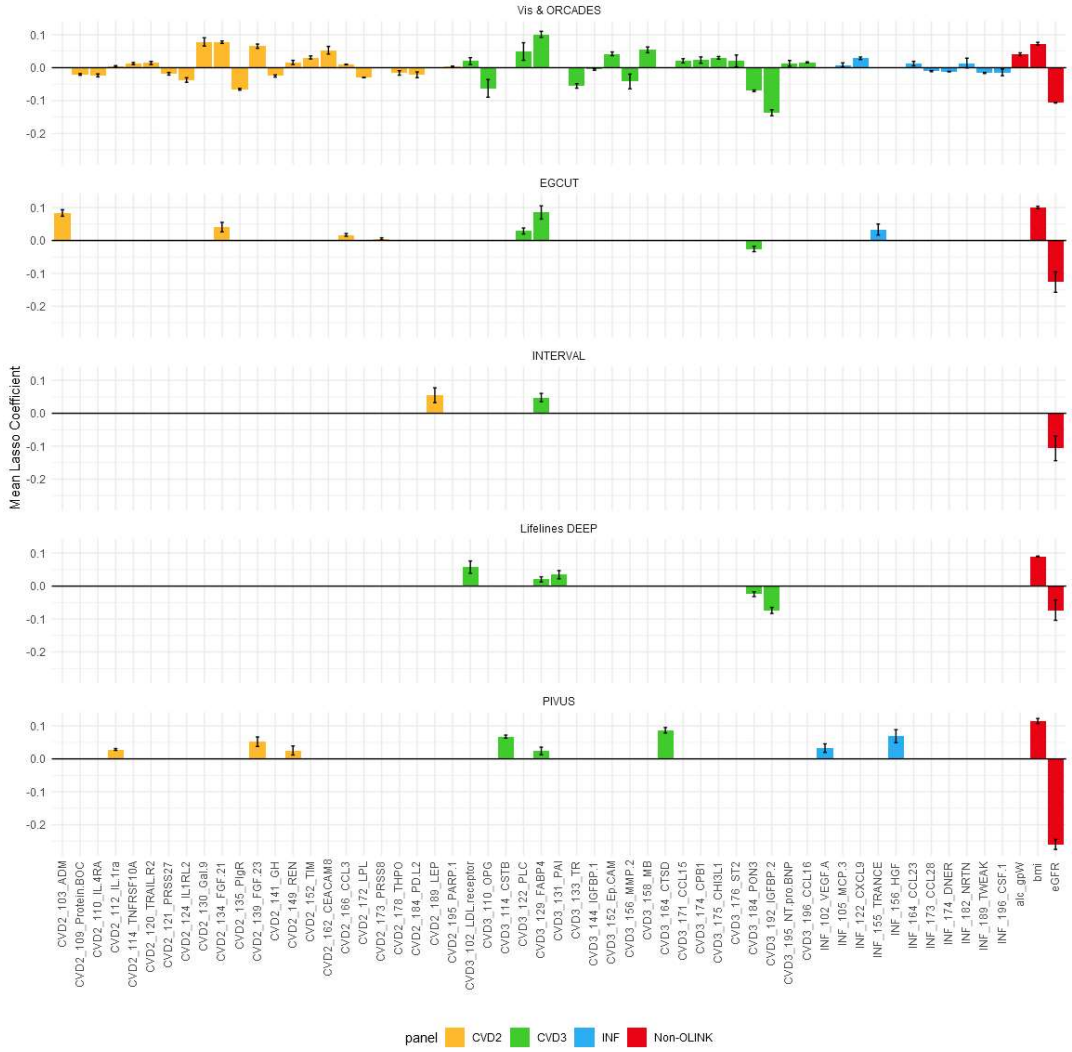


Figure 14 - Mean lasso coefficients for phenotypes included in >95% of models, per cohort.

Error bars are ±1 standard deviation.

2.3.5 Serum urate GWAS lookups

All proteins with significant partial correlations with serum urate or identified as urate-predictive in the lasso analysis in at least two cohorts were checked for shared genetic associations with serum urate. This set included FGF-23, Ep-CAM, IGFBP-2, FGF-21, CCL3, LDL-receptor, PLC, FABP4, REN, MMP-2, CHI3L1 and PON3. **Table 6** summarises which proteins were present in each lookup dataset and whether any pQTLs were identified.

Table 6 – pQTLs identified in each dataset

Protein	Vis & ORCADES	OLINK-IMPROVE	INTERVAL
CCL3	*Cis* & suggestive *trans*	*Cis* & t*rans*	*Cis*
CHI3L1	*Cis*	*Cis* & *trans*	*Cis*
EpCAM	*Trans*	N/A	N/A
FABP4	*Trans*	None	None
FGF21	*Trans* & suggestive *trans*	N/A	N/A
FGF23	*Trans*	None	None
IGFBP-2	*Trans*	N/A	None
LDL-receptor	None	N/A	None
MMP2	*Trans*	N/A	*Cis*
PLC	None	N/A	N/A
PON3	*Cis* & *trans*	N/A	N/A
REN	None	Trans	Cis

2.3.5.1 Vis & ORCADES

562 significant SNPs were identified and assigned to 10 loci across 9 proteins (**Figure 15**). The index SNPs are detailed in **Table 7**. No pQTLs were identified for PLC, LDL-receptor or REN. *Cis*-pQTLS were identified for CHI3L1, CCL3 and PON3; the remainder of signals were in *trans* to the gene encoding the protein, including one additional signal for PON3. None of the index SNPs were associated with serum urate using a Bonferroni-corrected p-value threshold of 0.05/10. One trans pQTL for EpCAM, at the FUT2 loci, associated with serum urate at a nominal level of significance.

When the p-value threshold was lowered to include suggestively-significant SNPs, a total of 1,873 were identified with p-values < $1x10^{-5}$ across all 11 proteins. These were assigned to 467 loci distributed across the genome. Of these, three had significant associations with serum urate after Bonferroni correction (P < 0.05 / 467). These SNPs are listed in the upper panel of **Table 8**. All three SNPs are in *trans* of the

associated protein. Two of these are associated with FGF-21, the third with CCL3, although the locus in the latter consists of a single rare variant.

The rare variant rs547836333, associated with CCL3, is imputed and present only in ORCADES. In the CCL3 meta-analysis, one ORCADES individual is heterozygous for the effect allele and 17 others have an imputed dosage between 0.15 and 0.65 copies. In the CKDGen meta-analysis, the SNP is present only in the Million Veteran Program (MVP) European subcohort – the variant was removed from the ORCADES results prior to meta-analysis due to the MAC > 10 filter. In ORCADES, the SNP is weakly negatively associated with associated with serum urate (-0.50, p = 0.049), but in the MVP it is strongly positively associated (5.56, $p = 8.58x10^{-153}$). Due to the very low allele frequencies and discrepancy of effect on serum urate, this genetic association can be assumed to be a false positive.

2.3.5.1.1 FGF-21 colocalisation

As the SNP rs780094 in the *GCKR* gene was significantly associated with serum urate and borderline significant with FGF-21 levels, a colocalisation analysis was performed for the 1Mb region centred on this SNP. These results are detailed in **Table 9**.

The strongest support was obtained for H_2, that the region contains a causal SNP for serum urate levels only (posterior probability = 0.83). However, the colocalisation test is sensitive to the sample size of the studies, and the *trans*-association for FGF-21 at the *GKCR* locus is only suggestively significant. The posterior probability of two separate causal SNPs was calculated as 0.16, and of one shared causal SNP at 0.004. With the available data, it appears to be more likely that the effects of the *GCKR* locus on serum urate and FGF-21 are driven by different variants. A larger GWAS of FGF-21 levels will confirm whether this *trans*-pQTL is real and allow clarification of the colocalisation result.

The region around the intergenic SNP rs799167 was also tested, and again the most likely scenario was that the region contained a causal SNP for serum urate levels but not for FGF-21 levels. (posterior probability = 0.820).

2.3.5.2 OLINK-IMPROVE CVD-I

The data available from the Olink IMPROVE website was found not to be in agreement with the results published in Folkersen *et al.*[115], in that I identified very different index SNPs in the online data from those in the paper. Additionally, CCL3

and FABP4 had in excess of 100 and 50 pQTLs respectively – this very large number is highly unlikely to be correct and calls the reliability of the online data into question. I have been in communication with the senior author to obtain updated results but have yet to acquire these at the time of writing.

As an interim solution, I have used index SNPs reported in the paper for CCL3, CHI3L1 and REN and looked these up directly in the serum urate meta-analysis results. Five index SNPs were looked up, two in CCL3 and two in CHI3L1, one in *cis* and one in *trans* in each protein, in addition to one *trans* pQTL for REN. None of these pQTLs were significantly associated with serum urate after Bonferroni correction ($P < 0.05 / 157$), but rs28601761 was associated at a nominal level ($P < 0.05$) - a *trans*-pQTL for CHI3L1 in *RP11-136O12.2*. These results are detailed in the middle panel of **Table 7**.

2.3.5.3 INTERVAL SOMAscan

Four of the eight proteins present in the SOMAscan data (FGF-23, CCL3, CHI3L1, FABP4, IGFBP-2, LDL-receptor, MMP2 and REN) had significant pQTLs. Three of these were in *cis*, for CCL3, CHI3L1 and REN, and the third was in *trans* of MMP2 at the *HSF2* locus. None of these loci were significantly associated with serum urate, either at a nominal ($P<0.05$) or a Bonferroni-corrected threshold. ($P < 0.05/4$). The results are detailed in the lower panel of **Table 7**.

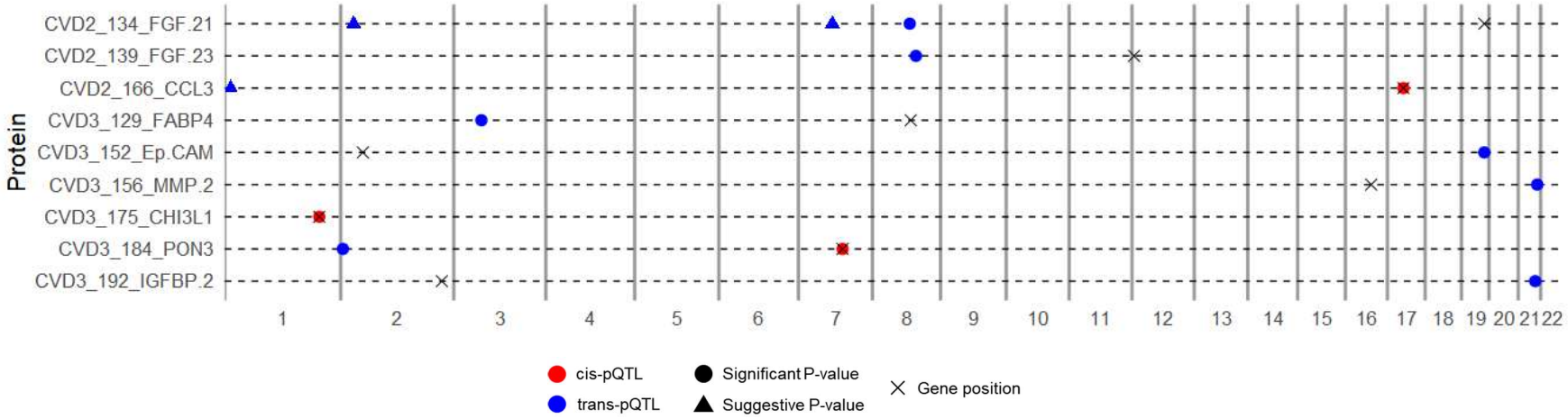


Figure 15 – Distribution of index SNPs from Olink meta-analyses of Vis & ORCADES.

Each track corresponds to one protein. X-axis corresponds to genome position. Each point represents an index SNP. Cis-pQTLs are plotted in red, trans-pQTLs in blue. SNPs marked with a circle (●) are significantly associated with levels of the corresponding protein (**Table 7**). SNPs marked with a triangle (▲) are suggestively associated with the protein but significantly associated with serum urate (**Table 8**). The location of the corresponding gene encoding the protein is given with a cross (x).

Table 7 – Lookup in CKDGen meta-analysis of serum urate (see Chapter 4) of pQTL Index SNPs from Vis & ORCADES, OLINK-IMPROVE and INTERVAL.

OLINK-IMPROVE Loci were taken from Table S1 in Folkersen et al[115]. Locus size data was not available for these SNPs. EAF from INTERVAL was also not available. SUA: serum uric acid. EAF: Effect allele frequency.

Protein	rsID	Chr.	Position	Locus size	Eff/Alt	EAF (Protein)	Effect (Protein)	P-value (Protein)	EAF (SUA)	Effect (SUA)	P-value (SUA)	Gene	Cis / Trans
Vis & ORCADES (Olink)													
CCL3	rs1719134	17	34,416,946	138	A/G	0.239	0.442	8.26E-33	0.265	-0.005	0.196	*CCL3*	Cis
CHI3L1	rs7556099	1	203,166,198	112	C/G	0.220	-0.623	8.76E-64	0.181	0.007	0.133	*CHI3L1*	Cis
Ep-CAM	rs570794	19	49,207,651	51	T/C	0.517	0.191	9.32E-09	0.509	-0.008	0.024	*FUT2*	Trans
FABP4	rs17060743	3	59,445,698	27	T/C	0.966	-0.531	6.92E-09	0.873	0.008	0.163	*RP11-719N22.2*	Trans
FGF-21	rs148345770	8	80,261,015	2	A/G	0.990	-1.038	1.40E-09	0.992	-0.018	0.438	*RP11-1114I9.1*	Trans
FGF-23	rs10504917	8	93,579,847	7	T/C	0.005	1.500	3.67E-09	0.254	-0.001	0.875	*RP11-587H10.2*	Trans
IGFBP-2	rs139234454	21	36,570,322	1	T/C	0.994	1.518	2.98E-08	0.983	-0.006	0.866	*RUNX1*	Trans
MMP2	rs183411833	21	40,733,749	7	T/C	0.013	-0.894	1.93E-08	0.010	-0.029	0.433	*PCP4*	Trans
PON3	rs773888624	2	4,865,648	1	A/G	0.006	-1.825	4.49E-08	0.005	0.008	0.958	*SNORA31*	Trans
PON3	rs149867961	7	95,025,744	216	T/C	0.947	0.736	9.06E-23	0.974	-0.007	0.589	*PON3*	Cis
OLINK-IMPROVE (Olink)													
CCL3	rs184243355	5	153,249,953	-	T/C	0.940	-0.41	2.2E-08	0.997	0.1271	0.527	*CTB-95D12.1*	Trans
CCL3	rs2188974	17	34,414,636	-	A/G	0.810	-0.310	4.90E-18	0.744	0.0046	0.210	*CCL3*	Cis
CHI3L1	rs2153101	1	203,168,474	-	A/T	0.210	-0.62	7E-108	0.181	0.0069	0.125	*CHI3L1*	Cis
CHI3L1	rs28601761	8	126,500,031	-	C/G	0.610	0.140	5E-09	0.375	-0.0104	0.002	*RP11-136O12.2*	Trans
REN	rs116661163	1	204,610,672	-	C/G	0.024	-0.718	1.03E-08	0.027	-0.0022	0.865	*LRRN2*	Trans
INTERVAL (SOMAscan)													
CCL3	rs712042	17	34,392,880	155	A/G	-	-0.714	1.35E-94	0.1801	-0.0019	0.695	*CCL18*	Cis
CHI3L1	rs884209	1	203,147,289	457	A/G	-	-1.013	~ 0	0.4629	-0.0022	0.498	*MYBPH*	Cis
MMP2	rs192645761	6	122,750,510	9	T/C	-	-0.573	4.79E-08	0.0148	0.008	0.697	*HSF2*	Trans
REN	rs193280350	1	204,148,649	5	A/G	-	1.199	1.32E-15	0.010	-0.0002	0.994	*REN*	Cis

Table 8 – Vis & ORCADES Index SNPs from suggestively significant loci ($P < 1 \times 10^{-5}$), with a significant association with serum urate ($P < 0.05/467$).

Protein	rsID	Chr.	Position	Locus Size	Eff / Alt	EAF (Protein)	Effect (Protein)	P-value (Protein)	EAF (SUA)	Effect (SUA)	P-value (SUA)	Gene	Cis / Trans
FGF-21	rs780094	2	27,741,237	14	T/C	0.391	0.183	5.77E-08	0.431	0.063	1.93E-87	*GCKR*	Trans
FGF-21	rs799167	7	73,051,306	35	T/C	0.703	0.185	3.59E-07	0.738	0.025	4.13E-12	Intergenic	Trans
CCL3	rs547836333	1	12,807,942	1	A/G	0.003	-2.994	5.90E-07	0.0002	5.556	8.58E-153	*C1orf158*	Trans

Table 9 – Colocalisation results for rs780094 and rs799167.

Phenotype 1 is serum FGF-21 level, phenotype 2 is serum urate level.

		rs780094			rs799167		
Hypothesis		**Prior**	**Bayes Factor**	**Posterior Probability**	**Prior**	**Bayes Factor**	**Posterior Probability**
H0	No association	0.547	4.51E-87	0.000	0.543	3.83E-47	0.000
H1	One variant associated with phenotype 1 only	0.186	2.60E-87	0.000	0.187	1.70E-47	0.000
H2	One variant associated with phenotype 2 only	0.186	1.00E+00	0.833	0.187	1.00E+00	0.820
H3	Two variants separately associated with phenotypes 1 and 2	0.063	5.76E-01	0.163	0.064	4.43E-01	0.125
H4	One variant associated with phenotypes 1 and 2	0.019	4.28E-02	0.004	0.019	6.68E-01	0.055

2.3.6 GENOSCORES

Correlations of greater than 0.1 were identified between 10 regions genetically associated with an Olink protein and 7 regions associated with serum urate. These are detailed in **Table 10**. Score-score correlations between these regions are shown in **Figure 16**. The strongest correlation identified was 0.79, between FGF-21 and serum urate in a region on the short arm of Chromosome 2. This region contains several genes, among which is glucokinase (hexokinase 4) regulator (*GCKR*), reported as the gene of interest in the GUGC serum urate GWAS. This protein was also identified as urate-predictive in the lasso regression (Section 2.3.4). FGF-21 was also weakly correlated (0.11) with a second region on Chromosome 2, in an intergenic region that had a negative correlation with MERTK.

The next strongest correlation is a negative correlation of -0.24 between CCL25 and serum urate on Chromosome 16, containing the genes *MIR1538*, *NFAT5* and *NQO1*. *NFAT5* was suggested as the likely causal gene in this region in the Köttgen *et al.* (2013) publication[74]. CCL25 had a nominally significant correlation with serum urate ($r = 0.050$, $P = 0.0462$) but was not significant in the partial correlation analysis, nor included in any lasso regression models.

No other correlations of magnitude larger than 0.2 were identified.

Table 10 – All GENOSCORES regions with a serum urate-Olink correlation > 0.1.

			SNP with minimum GWAS P-value				
	Region	No. of SNPs	RSID	GWAS P-value	GWAS Beta	Chr.	Position
Urate	Chr 2: 27,166,327 - 28,309,442	156	rs1260326	1.31×10^{-40}	0.0002	2	27,166,327
	Chr 2: 120,548,195 - 120,583,705	12	rs17050272	9.36×10^{-9}	0.0039	2	120,548,195
	Chr 5: 73,135,630 - 73,164,602	26	rs575416	3.63×10^{-9}	-0.0024	5	73,135,630
	Chr 6: 25,330,310 - 26,602,787	588	rs3799352	9.69×10^{-60}	-0.0005	6	25,330,310
	Chr 8: 23,857,479 - 23,930,882	28	rs17786744	8.82×10^{-8}	0.0014	8	23,857,479
	Chr 12: 110,953,995 - 112,468,611	13	rs653178	2.45×10^{-10}	-0.0046	12	110,953,995
	Chr 15: 75,866,642 - 76,020,780	13	rs1394125	$9.78 \times 10_{-11}$	0.0099	15	75,866,642
	Chr 16: 69,508,575 - 69,741,597	10	rs7193778	2.36×10^{-8}	-0.0037	16	69,508,575
Olink	**FGF-21**, Chr 2: 27,375,230 - 27,636,484	13	rs780094	5.77×10^{-8}	0.0234	2	27,375,230
	MERTK, Chr 2: 111,887,673 - 112,157,553	270	rs13386914	2.79×10^{-22}	0.0007	2	111,887,673
	ST1A1, Chr 5: 177,400,814 - 177,415,473	7	rs2545801	1.77×10^{-08}	0.0209	5	177,400,814
	IL-4RA, Chr 8: 20,304,455 - 20,380,215	23	rs2086882	6.64×10^{-8}	0.0007	8	20,304,455
	TNFRSF10A, Chr 8: 22,722,044 - 23,389,632	113	rs13278062	6.19×10^{-48}	0.0467	8	22,722,044
	EPHB4, Chr 12: 8,655,410 - 8,823,009	10	rs11047432	4.51×10^{-08}	-0.0071	12	8,655,410
	IL16, Chr 15: 81,237,069 - 81,825,241	48	rs4778639	6.15×10^{-101}	0.0404	15	81,237,069
	AP-N, Chr 15: 89,785,074 - 89,902,214	56	rs73478036	2.68×10^{-11}	-0.0019	15	89,785,074
	CCL25, Chr 16: 71,248,643 - 71,469,540	31	rs6499508	$3.81 \times 10_{-8}$	-0.0055	16	71,248,643

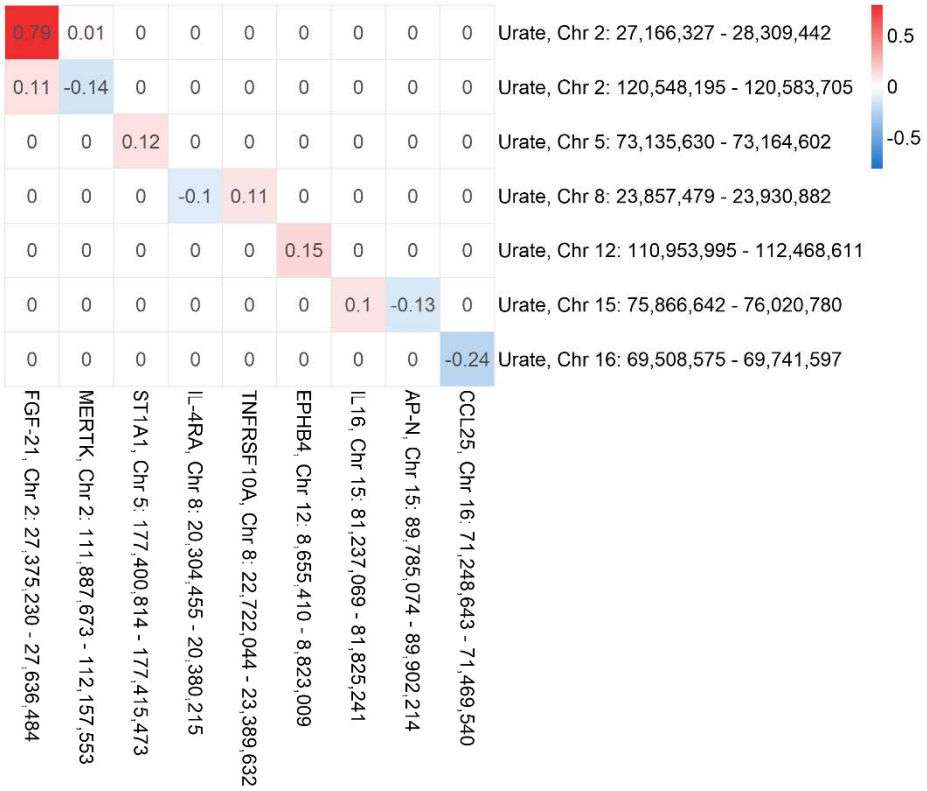


Figure 16 – GENOSCORES score-score correlation plot for serum urate and Olink loci.

Serum urate loci are plotted on the y-axis, Olink loci on the y axis. Correlation coefficient is given by the numerical value in each cell, as well as the colour-coding. Only loci with at least one correlation of absolute value greater than 0.1 with another locus are shown, for clarity.

2.3.7 Genetic correlation with LD-score regression

Genetic correlations are shown in **Table 11**. No traits had significant genetic correlations. Traits could not be calculated as the LDSC algorithms are unstable for samples with n < 3,000. FGF-23 has a near-significant P-value for a correlation of 0.07 in the Vis + ORCADES dataset, but although the correlation is stronger using OLINK-IMPROVE data, the standard error is also much larger, resulting in a non-significant correlation. This may be because the sample size is still below the recommended minimum of 3,000 but may also be a consequence of the problems identified earlier with the IMPROVE data. The heritability of the OLINK traits would also affect the genetic correlation; unfortunately, these have not yet been calculated to my knowledge. Rerunning this analysis when the SCALLOP CVD-II, CVD-III and

INF results are published should clarify this. FABP4 displayed the strongest genetic correlation with serum urate, 0.30 (P = 0.09).

Table 11 – Serum urate-Olink genetic correlations from LDSC.

Missing values are from algorithm failure due to low sample size.

	Protein	Genetic correlation r_g	SE	P-value
Vis + ORCADES	FGF-23	0.0727	0.0405	0.0724
	CCL3	-0.0428	0.0691	0.5355
	LDL-receptor	-	-	-
	PLC	-	-	-
	FABP4	-	-	-
	Ep-CAM	0.0644	0.0913	0.4805
	MMP.2	-0.0016	0.0519	0.9762
	CHI3L1	0.0121	0.0312	0.6992
	PON3	0.0013	0.0449	0.9777
	IGFBP-2	-	-	-
IMROVE	CCL3	0.2017	0.3847	0.6001
	CHI3L1	-	-	-
	FABP4	0.3016	0.1783	0.0907
	FGF-23	0.2244	0.3952	0.5702

2.4 Lipidomics results

2.4.1 Analysis sample sizes

The samples sizes for each cohort and sub-analysis are displayed in **Table 12**. These are consistent across all subsequent analyses.

Table 12 – Sample sizes for serum urate-lipidomic correlation analyses.

Cohort	Sub-analysis	N
CROATIA-Vis	All	662
	CKD-excluded	613
	Females only	390
	Males only	272
ORCADES	All	577
	CKD-excluded	563
	Females only	294
	Males only	283
Merged	All	1239
	CKD-excluded	1176
	Females only	684
	Males only	555

2.4.2 Correlations

Correlations between serum urate and lipidomic measurements are shown in **Figure 17**. A total of 99 lipidomic phenotypes were significantly correlated with serum urate after FDR correction for multiple testing for 191 tests. Additionally, 4 of the non-omic phenotypes included were significant, including HDL and LDL cholesterol, which were not included in the Olink analysis. The strongest absolute correlation was again with BMI, and the strongest lipidomic correlation was with phosphatidylethanolamine 40:6 (PE_40__6, r = 0.267). The strongest negative correlation was with phosphatidylcholine O-34:2 (PC_O_34__2, r = 0.180).

There were three cases where a significant difference in effect between sexes was detected. These are detailed in **Table 13**. All of these are phosphatidylcholines, two of which have negative correlations in females and positive in males, and the third has a strong positive correlation in males and a near-zero correlation in females.

Table 13 – Serum urate-lipidomic correlations with a significant difference in effect between sexes.

PC_38__6 – Phosphatidylcholine 38:6; PC_38__7 – Phosphatidylcholine 38:7; PC_O_40__6 - Phosphatidylcholine O-40:6, n_F = 906, n_m = 710.

Phenotype	r_F	r_M	$P_{sex\ effect}$
PC_38__6	0.005	0.200	8.27E-05
PC_38__7	-0.084	0.101	2.16E-04
PC_O_40__6	-0.157	0.034	1.31E-04

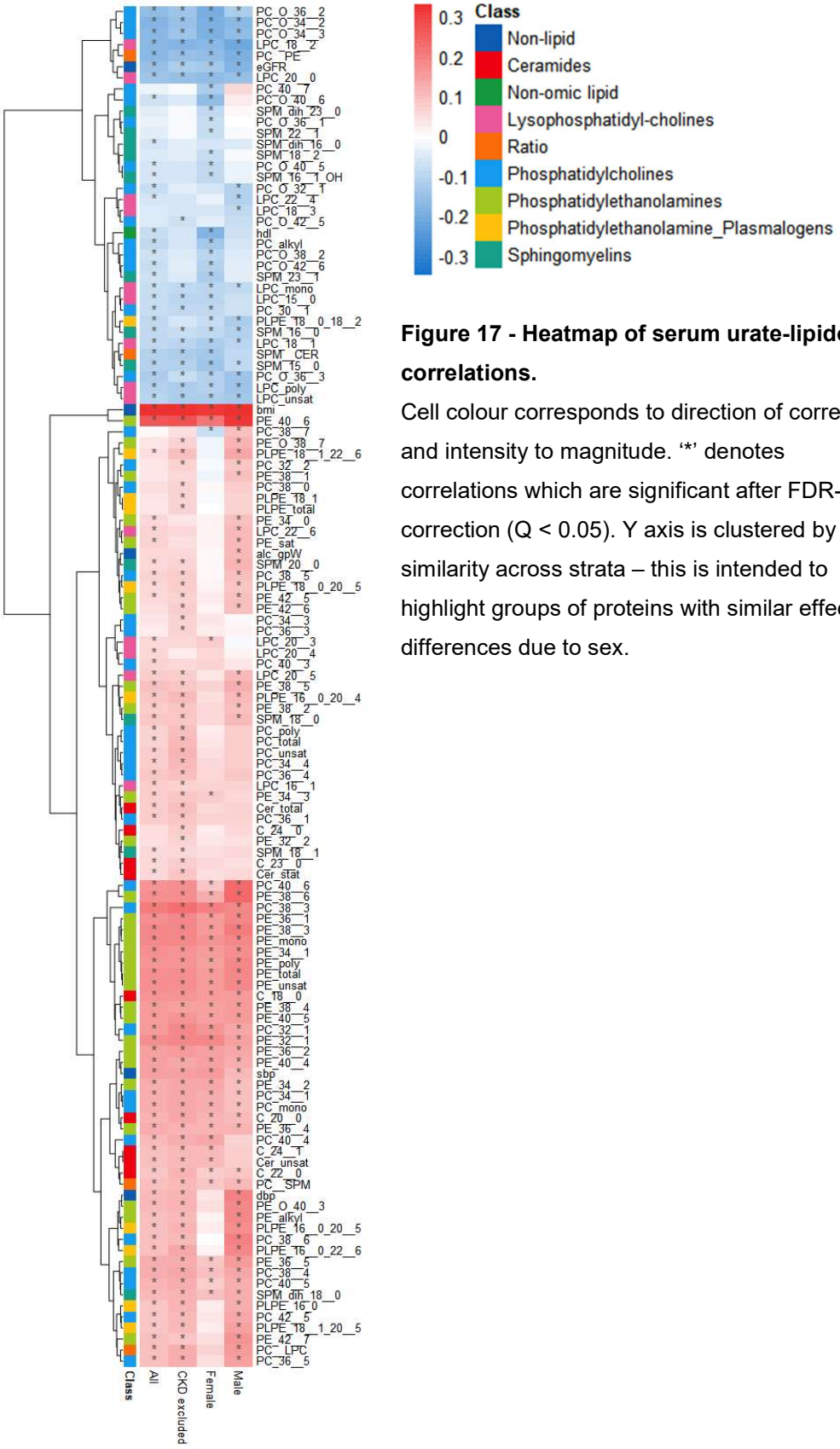


Figure 17 - Heatmap of serum urate-lipidomic correlations.

Cell colour corresponds to direction of correlation and intensity to magnitude. '*' denotes correlations which are significant after FDR-correction (Q < 0.05). Y axis is clustered by similarity across strata – this is intended to highlight groups of proteins with similar effect differences due to sex.

2.4.3 Partial Correlations

Three phenotypes were significant in the partial correlation analysis (**Table 14**, **Figure 18**). In contrast to the Olink partial correlation analysis (Section 2.3.3) two of these were non-omic phenotypes. The third, phosphatidylcholine 38:6 (PC_38__6), is negatively correlated with serum urate levels in all four subsets of the data, in contrast to the results observed in normal correlation analysis, where all correlations were positive or zero, and a significant difference in effects was observed between males and females.

Table 14 – Partial correlation coefficients for BMI and eGFR in serum urate-lipid models. Rows in gold are significant after FDR correction.

Phenotype	Subset	Partial correlation coefficient	Std. Err.	P-value	Q-value
BMI	All	0.128	0.026	3.41E-05	2.17E-03
BMI	CKD excluded	0.139	0.031	1.17E-05	1.12E-03
BMI	Female	0.122	0.037	6.66E-03	3.24E-01
BMI	Male	0.102	0.043	5.13E-02	6.70E-01
eGFR	All	-0.186	0.026	1.34E-09	1.28E-07
eGFR	CKD excluded	-0.129	0.031	4.65E-05	2.96E-03
eGFR	Female	-0.139	0.037	1.94E-03	1.85E-01
eGFR	Male	-0.235	0.043	5.58E-06	5.33E-04
PC_38__6	All	-0.096	0.026	1.82E-03	8.68E-02
PC_38__6	CKD excluded	-0.109	0.031	6.15E-04	2.94E-02
PC_38__6	Female	-0.119	0.037	8.20E-03	3.24E-01
PC_38__6	Male	-0.090	0.043	8.41E-02	6.70E-01

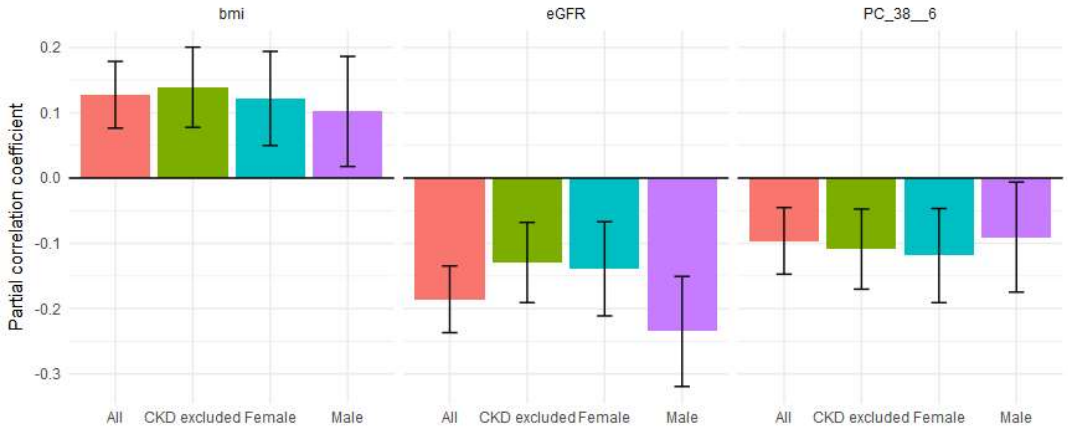


Figure 18 - Partial correlation coefficients for BMI and eGFR in serum urate-lipid models.

Error bars are 95% confidence interval.

2.4.4 Lasso regression analysis

Phenotypes were considered urate-predictive if they appeared in more than 950 models. Results are displayed in **Figure 19**.

The phenotypes with the strongest contribution to the models were BMI, eGFR and phosphatidylethanolamine 40:6 (PE_40__6). No other phenotypes had coefficients with magnitude larger than 0.1. Coefficient sign was consistent for lysophosphatidylcholines (LPC_x__y, negative), acyl-group phosphatidylcholines (PC_O_x__y, negative) and phosphatidylethanolamines (PE_x__y, positive). No consistent direction of effect was seen between sphingomyelins or between phosphatidylcholines lacking ether links (PC_x__y).

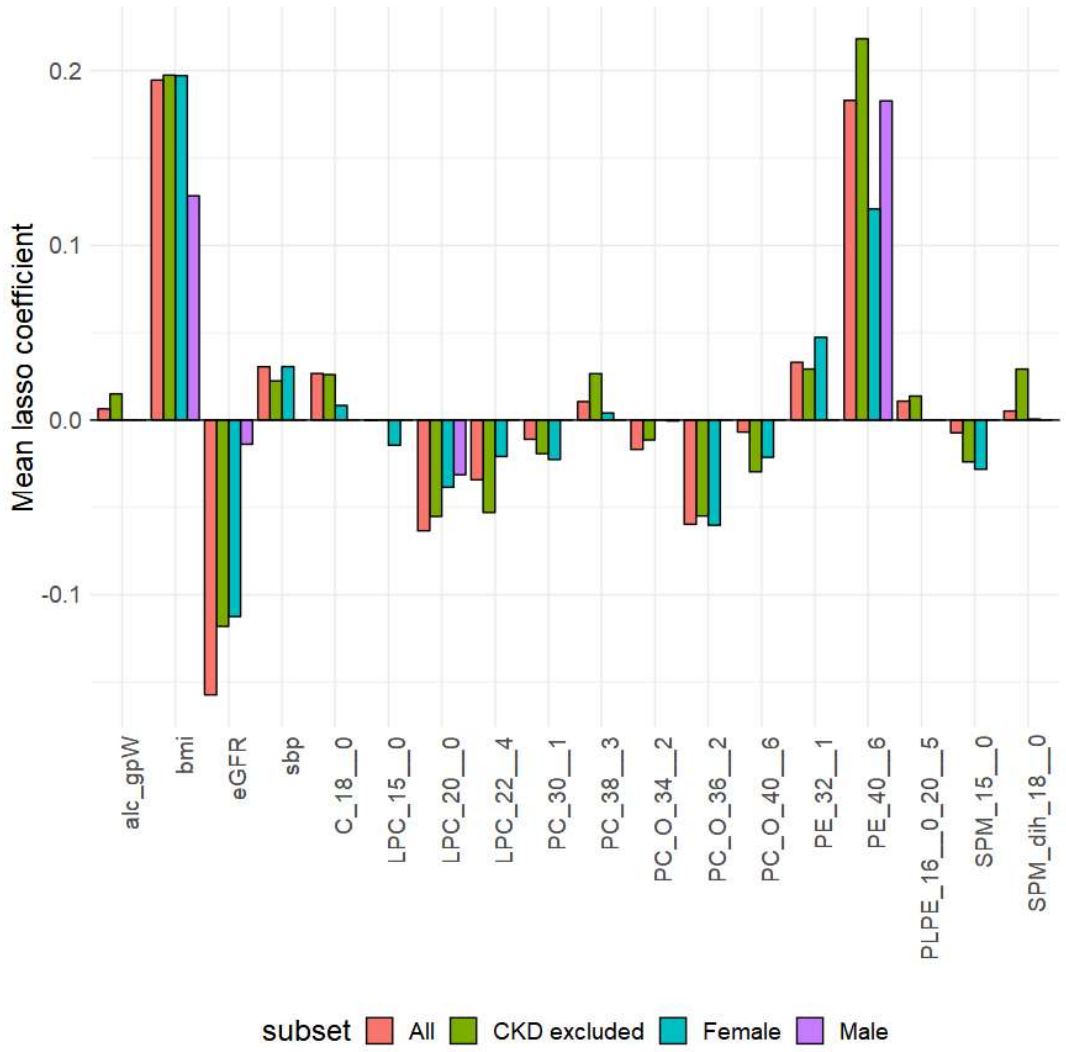


Figure 19 – Mean lasso regression coefficients for serum urate-lipidomics analysis. Means are calculated over 1000 runs. All phenotypes appearing in 95% of models in at least one subset are shown.

2.5 Combined analysis

Data from both Olink and lipidomics were combined to investigate the overlap between these sets of measurements. Pairwise correlation analyses are independent of other phenotypes, hence these would not be changed in the merged analysis, so these were not re-run. Partial correlation and lasso regression analyses are both sensitive to covariates included, so these were performed *de novo* using the combined dataset.

No phenotypes were significantly partially correlated, although the strongest correlation, with FGF-23 (r = 0.131, P = 2.42E-4, Q = 0.055), was very close to the threshold of significance.

The lasso regression includes many of the phenotypes observed in the separate analyses, including the strong associations with eGFR, IGFBP-2, PON3, LPC 22:4, FGF-23, FGF-21, Gal9, BMI, PE 40:6 and FABP-4. These results are shown in **Figure 20**.

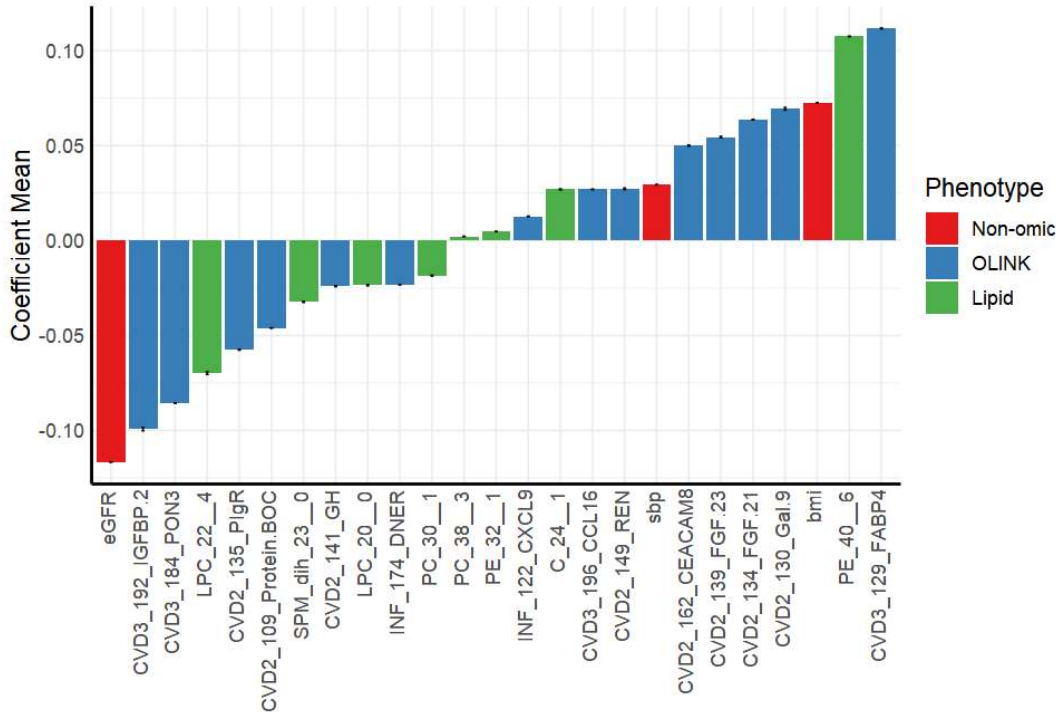


Figure 20 - Mean lasso coefficients for serum urate regressed on all phenotypes. Only phenotypes included in all 1000 iterations of the model are plotted.

2.6 Conclusions

The analyses in this chapter, though broadly similar in theme, identify a range of different circulating proteins and lipids as being associated with serum urate. Correlation analysis identifies a large number of associations, many of which are likely attributable to confounding with phenotypes such as BMI or eGFR, which are known to be correlated with serum urate. Partial correlation analysis identifies the fewest associations but is restrictive in that it fixes the effect of all covariates and cannot

account for combined effects of multiple phenotypes. Lasso regression attempts to find a small set of phenotypes whose combination of values best predicts serum urate levels, in this way, it is less restrictive than partial correlations, as it can allow more than one phenotype to be assessed at a time, but it is less subject to confounding than basic correlations, as redundant phenotypes add little information to the model and are pruned. One limitation of the lasso method is that it enforces sparsity. In a biological context, this may not necessarily reflect the mechanisms regulating a relationship, which are often multiple and complex. Partial correlations allow the analyst to consider each protein individually in the context of any other phenotypes of interest – a non-significant correlation may still be of interest – while lasso regression simply sets some coefficients to zero.

All these analyses detect known associations between serum urate and protein levels, suggesting that they are at the very least technically correct in their execution. More interestingly, several phenotypes are identified which have not been previously reported as correlated with serum urate. Though many studies have investigated the relationship between serum urate and a limited number of phenotypes, to my knowledge this dataset represents the broadest study of urate-phenotype associations, particularly with respect to cardiovascular- and inflammatory-related omic phenotypes.

In addition, the genetic architectures of molecular phenotypes are often simple, with a few variants conferring strong effects on abundance[151–153] – for example, six pQTLs collectively explain up to 52% of the variance in VEGF levels[154]. These variants can be used to interrogate relationships; **Figure 21** summarises the possible mechanisms for a genetic variant affecting both biomarker and serum urate levels. Of particular interest is the middle scenario: where a variant is both a QTL for serum urate and a strong *cis*-pQTL for the variant of interest. In this case, it is likely that the change in protein levels is not caused by variation in serum urate, rather that the change in serum urate levels is mediated by the altered protein levels. If the variant is not pleiotropic, it can be used as an instrumental variable in Mendelian randomisation analysis to formally test the causality in the relationship.

Unfortunately, none of the *cis*-pQTLs identified or previously reported for the proteins selected as urate-associated showed any association with serum urate, even at a nominal significance level ($p < 0.05$). This could be due to the alternative hypothesis that serum urate levels influence the abundance of some of these protein biomarkers.

Alternatively, as the GENOSCORES data suggests in some cases, some of the associations may be independent consequences of the same underlying pathway. For example, *GCKR* is a QTL for serum urate and a borderline significant *trans*-pQTL for FGF-21 (P ~ $5x10^{-8}$), suggesting they may be independent consequences of the glucokinase regulation pathway. Two *trans*-pQTLs are nominally significant with serum urate levels, *FUT2* for Ep-CAM in our data (p = 0.024) and *RP11-136O12.2* for CHI3L1 in IMPROVE (p = 0.002) which could mean the *trans*-genes have independent regulatory effects on both serum urate and the protein.

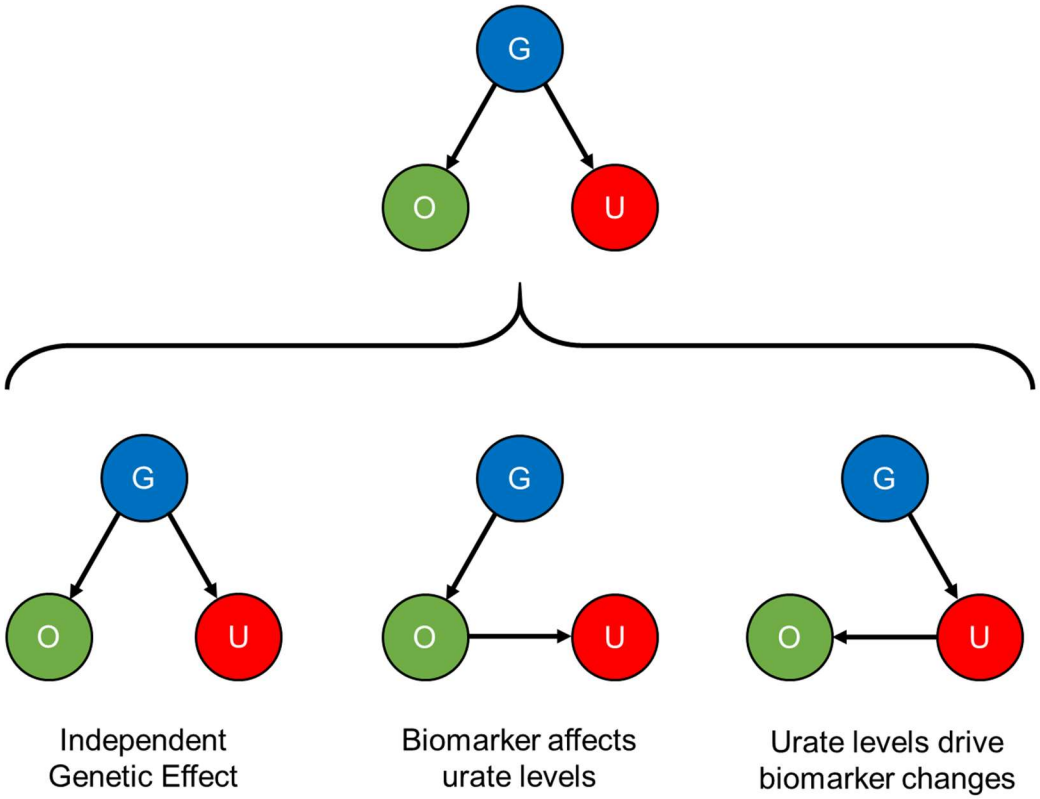


Figure 21 - Possible mechanisms driving an association between serum urate (U) and omic biomarker (O) levels where a shared genetic variant (G) can be identified.

2.6.1 FGF-23

FGF-23 was identified as positively associated with serum urate in both partial correlation and lasso regression analyses. It is part of a subfamily of fibroblast growth factors that includes FGF-19, 21 and 23, which are noteworthy within the FGF family for being endocrine hormones – the other five FGF subfamilies operate in a paracrine manner, affecting mainly the tissues from which they are secreted[155]. FGF-21 was

also implicated in these analyses as being linked to serum urate, though FGF-19 was not.

FGF-23 is a hormonal regulator of phosphorus and vitamin D metabolism, most highly expressed in bone cells. Amongst other known roles, it regulates reabsorption of phosphate in the kidneys by modulating sodium phosphate co-transporters. Glycosylated FGF-23 binds to the membrane-spanning receptor protein FGFr1 in the kidney, causing a signalling cascade that reduces the reabsorption of phosphate into the blood. FGF-23 also downregulates enzymes which activate vitamin D. Because activated vitamin D enhances intestinal absorption on phosphate, this means FGF-23 decreases phosphate levels though an additional pathway[155].

FGF-23 is increased in cases of non-alcoholic fatty liver disease (NAFLD), independently of low vitamin D[156] – a condition also associated with hyperuricaemia in both East Asian and European populations[157,158]. Elevated FGF-23 levels are observed in CKD, but the persistence of this correlation in the CKD-excluded analyses suggests that the association goes beyond impaired renal function. FGF-23 protein and mRNA levels were found to be increased in the kidney of a CKD model rat[159], and is a strong predictor of insulin resistance in CKD patients[160].

In a rat hepatocyte model using a uricase inhibitor, increased uric acid production was found to be an indicator of com promised ATP synthesis[14]. As urate is the end product of AMP oxidation, this may suggest that the link between serum urate and FGF-23 is driven by phosphate homeostasis – high FGF-23 leads to low phosphorus levels, causing reduced ATP synthesis and more AMP build-up, which is consequently oxidised to uric acid.

Positive associations between FGF-23 and uric acid have been reported in healthy adult males[161], CKD cases[162] and children with kidney damage but preserved kidney function[163]. These studies incorporate various covariates, including BMI and eGFR, but none have adjusted for such a large number and range of phenotypes. The fact that the association persists even after the adjustment for so many medically-relevant phenotypes is good evidence that the association between FGF-23 and serum urate is genuine.

Unfortunately, the absence of a *cis*-pQTL for FGF-23 in our data makes it difficult to test the causality of the association. None of the *trans*-pQTLs identified were associated with serum urate, even at a nominal threshold of $p < 0.05$, making it

unlikely that any of these genes are driving the correlation between FGF-23 and serum urate.

2.6.2 FGF-21

FGF-21 was positively predictive of serum urate in the lasso regression. It is a mediator of the fasting response that increases glucose uptake and increases insulin sensitivity, and is expressed in the liver, thymus, adipose tissue and islet β-cells in the pancreas. Elevated levels have been linked to insulin resistance, metabolic syndrome and NAFLD. The mechanism by which it increases glucose uptake is partly due to upregulating transcription of GLUT1[164], a glucose transporter in the same family as the high-capacity urate transporter GLUT9, encoded by *SLC2A9*. This regulatory activity means FGF-21 has a beneficent impact on type 2 diabetes[165].

FGF-21 has been previously positively correlated with uric acid levels, adjusted for age and BMI in a small study of 210 Mexican ancestry individuals, where it was also found to correlate with fasting glucose and physical activity[166], but to my knowledge has not been reported elsewhere. Cuevas-Ramos *et al.* suggest that the association between metabolic syndrome and FGF-21 may explain the link with serum urate, but their associations persist even after excluding metabolic syndrome cases.

GCKR encodes glucokinase regulatory protein (GKRP), a hepatocyte-specific inhibitor of glucokinase, one of the enzymes which convert glucose to glucose-6-phosphate. A mutation in *GCKR* has been reported to affect expression of FGF-21[167]. This SNP, rs1260326 (P446L), has been well characterised – *in vitro* experiments showed that it results in elevated hepatic glucose uptake and disposal, which increases lipid synthesis and reduces fasting glucose[168]. In GWAS, it has been associated with reduced T2D risk[169], reduced fasting plasma glucose and increased triglyceride and CRP levels[170].

GCKR has also been associated with serum urate levels[171]. This is borne out by the strong GENOSCORES correlation near GCKR between FGF-21 and serum urate. A genetic association between the two traits supports a hypothesis that increased glucokinase activity leads to more ATP depletion and thus higher serum urate levels. Unfortunately, the pleiotropy of the *GCKR* locus means it would not be an ideal instrument for MR.

2.6.3 IGFBP-2

IGFBP-2 was consistently negatively associated with serum urate across the analyses, particularly strongly in the discovery cohort. Association strength was affected, although it remained suggestive, after removal of CKD cases, in line with IGFBP-2 as a biomarker of renal pathology in CKD and lupus nephritis patients[172] .

This protein modulates the activity of Insulin-Like Growth Factors, binding IGF-I and IGF-II in the blood and many different ligands inside the cell. IGFBP-2 has been linked to modulation of metabolism, with lower levels associated with diabetes[173], obesity[174] and metabolic syndrome[175], as well as bone development in mice, directly stimulating osteoblast differentiation[176]. A GWAS hit for type 2 diabetes has been found in an intron of the *IGFBP2* gene[169]. Additionally, high expression of the *IGFBP2* gene has been shown to promote the growth of several types of tumours.

In a study of changes of protein abundance in obese individuals following a low-calorie diet by Carayol *et al.*, IGFBP-2 levels were strongly negatively associated with BMI at baseline, but this effect was greatly reduced in magnitude following intervention[117]. This is consistent with my finding that IGFBP-2 is negatively correlated with serum urate, which is increased in obesity. However, the fact that IGFBP-2 was retained in the partial correlation and lasso predictive models despite the inclusion of BMI as a separate covariate suggests that the link between serum urate and IGFBP-2 is not simply mediated by BMI.

A single *trans*-pQTL for IGFBP-2 was identified in the *RUNX1* locus, but this was not associated with serum urate. *RUNX1* encodes a transcription factor that plays a role in the differentiation of blood cells.

2.6.4 Ep-CAM

Ep-CAM was identified in the discovery partial correlation analysis and had consistent positive correlation in INTERVAL but did not replicate in EGCUT or Lifelines DEEP. It was not identified as urate-predictive in the lasso regression analysis. The protein is a transmembrane mediator of cell-cell adhesion in epithelial tissue and has been reported as an oncogene. It is strongly expressed in glandular cells of the gastrointestinal tracts and the gall bladder, as well as in the cells of the kidney tubules[177] – which are composed of epithelial tissue – but otherwise there is no obvious connection to serum urate levels.

A cis-pQTL was identified at rs201314303 by Enroth *et al.*[178], but this SNP was not present in the CKDGen meta-analysis, and the nearest index SNPs were several megabases away. In our own meta-analysis, rs570794, a *trans*-pQTL in the *FUT2* gene was nominally significant with serum urate ($p < 0.05$).

FUT2 encodes a fucosyltransferase involved in ABO antigen synthesis[179]. The SNP is also associated with alkaline phosphatase[180], alcohol intake frequency[140] and self-reported high cholesterol[140] in the UK Biobank. A pleiotropic association with alcohol could explain the serum urate association, but the relatively high p-value with serum urate makes this link tenuous.

2.6.5 Renin

Renin (REN) was included in lasso models for the discovery and PIVUS cohorts as a positive predictor of serum urate levels. Renin is a regulator of blood pressure as part of the renin-angiotensin system (RAS), where it converts angiotensinogen to angiotensin I, which is subsequently converted to angiotensin II, a vasoconstrictor. Uric acid is known to be positively correlated with systolic and diastolic blood pressure[181] and has been linked to the RAS. Transcript levels of the *REN* gene, which encodes renin, along with other components of the RAS pathway were shown to increase in a dose-dependent manner in differentiating mouse 3T3-L1 adipocytes incubated with uric acid[182]. This study goes on to suggest that RAS may be the mechanism by which uric acid is linked to obesity-related hypertension. Renin is also positively correlated with blood pressure[183], but in the combined Olink and lipids analysis, where systolic blood pressure (SBP) was included as a covariate, both SBP and renin are identified as independent predictors of serum urate levels. This suggests that the relationship between serum urate and renin is not simply mediated by high blood pressure. The Carayol *et al.* obesity dietary intervention study did not find an association between REN and BMI or BMI decrease, suggesting that the serum urate association is unlikely to be mediated by BMI[117]. However, it has been suggested[8] that serum urate could be increased due to dysregulation of the RAS in obesity – increased angiotensin II reduces fractional clearance of uric acid (the ratio of uric acid clearance to creatinine clearance) in the kidney[184] and thus increased uric acid in the bloodstream.

No pQTLs for REN were present in in our own meta-GWAS. A *cis*-pQTL was identified in INTERVAL, but it was not significantly associated with serum urate. Additionally, a

trans-pQTL was reported by Folkersen *et al.*[115]at rs116661163 in the *LRRN2* locus, which was not significant with serum urate (P > 0.05)

2.6.6 LDL-receptor

LDL-receptor was included in lasso regression models in the discovery and Lifelines DEEP cohorts. This protein is a cell surface receptor involved in receptor-mediated LDL endocytosis. Loss of function mutations of LDLR cause familial hypercholesterolaemia. LDL-cholesterol is positively correlated with serum urate levels[185], which may explain the correlation with LDLR. As far as I can ascertain, no association between LDLR and serum urate has been reported in humans.

In rabbits, LDLR protein was found to be expressed at a lower level in the liver of diet-induced hyperuricaemic rabbits compared to controls, but its level restored to some extent following treatment with Losartan[186]. The hypertension drug Losartan is an angiotensin receptor blocker with a known effect on reducing serum urate level. Losartan directly inhibits URAT1[187], a urate/anion exchanger encoded by *SLC22A12*. This transporter is one of several present on the apical membrane of the kidney proximal tubule cells, where it reabsorbs urate from the tubule lumen[37]. Losartan blocks this reabsorption, leading to uricosuria – high levels of uric acid in the urine. However, it must be noted that serum urate levels are much lower in rabbits than humans. Additionally, although it is commonly noted that Losartan has no effect on cholesterol levels, it cannot be completely ruled out that this is driving the effect on LDLR in the rabbit model.

There is no pQTL for LDL-receptor either in our own data or in the literature. However, the multiple GWAS signals locating to the *LDLR* gene have no association with serum urate levels, while still being strongly associated with lipid levels[188]. Taken together, this evidence strongly suggests there is no causal link between LDLR and serum urate. However, with the available data, I cannot exclude a pleiotropic effect on both LDLR and serum urate, for example mediated by LDL levels, or serum urate level causally affecting LDLR.

2.6.7 FABP4

FABP4 was one of only two phenotypes that was included in all Olink lasso models for all five cohorts – the other being eGFR, a known strong association with serum urate. This association persists in the combined Olink-lipids lasso analysis (**Figure**

20) and is assigned the largest positive coefficient of any phenotype. FABP4, a fatty acid binding protein also known as aP2, is primarily expressed in adipocytes and also in macrophages, though at a 10,000-fold lower level[189]. Adipose tissue has been proposed to be a site of urate production in obese individuals based on work in mice[190]. The dual metabolic and inflammatory role of FABP4 mirrors the role of uric acid and suggests a link to diabetes and metabolic syndrome.

I have found little evidence of the link between serum FABP4 and serum urate being explicitly stated before – a Chinese paper from 2010 appears to have reported a correlation between serum FABP4 levels and serum urate in pregnant women with preeclampsia, but the precise methodology is unclear to me, as only the abstract is in English[191]. A study in mouse 3T3-L1 preadipocytes found that the uricosuric drug benzbromarone, reduces levels of human FABP4, but the authors deduce that this is due to a direct interaction, making no mention of any association between FABP4 and serum urate[192]. Benzbromarone acts as a non-competitive inhibitor of XO, the enzyme which catalyses the conversion of hypoxanthine to xanthine, and xanthine to uric acid. However, the effects of XO inhibitors are often difficult to interpret, as they reduce the production of reactive oxygen species as well as serum urate.

Another study using differentiating 3T3-L1 cells found that incubation with benzbromarone increased levels of *FABP* mRNA (referred to as *aP2* in the study), although this response was reduced in the cells that had been differentiating for longer. Allopurinol and uric acid had no effect on *FABP4* transcript abundance[193]. This would seem to suggest that serum urate is not directly stimulating *FABP4* transcription under these conditions, although the effect could again be driven by reduced ROS production. It must also be noted that *in vitro* experiments on mouse cells do not necessarily represent the *in vivo* systems in humans.

A trans-pQTL in *RP11-719N22.2* in our meta-analysis was not significant in serum urate. No information on FABP4 was reported in either the IMPROVE or INTERVAL studies, but it will be included in the next round of SCALLOP meta-analysis results.

2.6.8 IL-1RA

IL-1RA is not strongly predictive of serum urate levels in the discovery cohort but is retained in all models. It is more strongly predictive in PIVUS, which only included a reduced set of Olink phenotypes. The regression coefficient is positive in both, which contradicts a published finding that peripheral blood mononuclear cells (PBMCs)

stimulated with urate showed a downregulation of *IL-1ra* transcription[194], although this study used uric acid levels that were much higher than those seen in clinical hyperuricaemia. This could also be an effect of the inclusion of many covariates in my models.

2.6.9 CCL3

C-C Motif Chemokine Ligand 3 (CCL3) is a small inducible cytokine which was a weak association in the lasso regression for discovery and EGCUT. CCL3 has been shown to be produced by eosinophils in response to uric acid crystal stimulation[195].

A suggestive trans-pQTL ($P < 10^{-5}$) is significantly associated with serum urate in the *C1orf158* gene, which is uncharacterised. However, as detailed in section 2.3.5.1, this SNP is very rare, and is present only in ORCADES in the CCL3 meta-analysis and the MVP cohort in the serum urate meta-analysis, with opposite effects on serum urate, so, unfortunately, it is likely to be a false positive.

2.6.10 PLC

Perlecan (PLC) was identified in the lasso regression in the discovery and EGCUT cohorts. It is a large multidomain proteoglycan that cross-links extracellular matrix components and cell-surface molecules and is a major component of the glomerular basement membrane. A study in canine kidney cells found that 8mg/dL of uric acid caused decreased expression of the perlecan core protein, but not 20mg/dL[196]. The lasso regression shows a mild positive association, but mechanism behind the possible link with serum urate is not clear. No pQTLs were identified.

2.6.11 PON3

Paraoxonase 3 (PON3) is associated with HDL-cholesterol in the blood, and reduces the rate of LDL-cholesterol oxidation, low levels of which have been linked to a variety of conditions including inflammation in coronary artery disease (CAD) and insulin sensitivity in peripheral artery disease[197]. It is mainly expressed in the kidney and liver.

PON3 was identified as urate-predictive in three of the four cohorts in which it was measured, absent only from INTERVAL. In all three cohorts it was negatively associated with serum urate levels.

No previous associations with serum urate have been reported in the literature, but a strong negative correlation was reported between the activity of related protein PON1

and uric acid levels in the blood of patients undergoing haemodialysis[198]. This paper suggests that PON1 acts to protect cholesterol from oxidative damage which is less necessary at higher concentrations of serum urate.

A *cis*-pQTL and a *trans*-pQTL in *SNORA31* were identified in our meta-analysis, but neither were significantly associated with serum urate.

2.6.12 Lipid associations

Interpretation of the lipidomic results is more complicated than the Olink, as the distinction between different lipids is less, and each has been studied far less than the disease biomarkers included on the protein panels. However, there are still a few notable phenotypic correlations.

By far the strongest association between serum urate and a lipid was with phosphatidylethanolamine 40:6 (PE 40:6). Phosphatidylethanolamines (PEs) comprise 25% of all phospholipids in the body. It is not clear why PE 40:6 displays such a strong positive correlation with serum urate compared to the other PEs measured (**Figure 17**), but it persists in lasso regression in both lipids and the combined sample, and so can be considered to be an independent predictor of serum urate levels.

Phosphatidylcholine 38:6 was the only lipid detected in the partial correlation analysis. The identified correlation was negative, which reflects previous findings that decreases in PC 38:6 are associated with progression from cognitive impairment to Alzheimer's disease[199] and development of Huntington's disease[200], both of which are associated with hypouricaemia. No association was detected in the lasso analysis. Interestingly, PC 38:6 is positively correlated with serum urate in the correlation analysis, but negatively partially correlated. This could mean that one of the covariates adjusted for accounted entirely for the positive correlation seen between serum urate and PC 38:6, revealing the true association.

PC 38:6 is also notable for its strong sex effect: it is positively correlated with serum urate in males, but no correlation is detected in females. Sex effects were also detected in PC 38:7, PC 40:7 and PC O-40:6, in all cases the association was negative in females and positive in males.

Lysophosphatidylcholine 22:4 has a moderate negative effect in the combined lasso, although the effect in the lipids-only lasso is comparatively weak.

Lysophosphatidylcholines (LPCs) are produced from PCs through the removal of one of the fatty acyl groups. They are minor components of the cell membrane but are present at higher levels in plasma and are major components of oxidised LDL. Most LPCs were negatively correlated with serum urate, and all the LPCs retained in regression models had negative coefficients. The PC:LPC ratio is altered in many different diseases, but I was unable to find a consistent pattern that might explain the negative associations with serum urate, or a record of a specific role for LPC 22:4.

2.6.13 Summary

The association of several medically-relevant proteins with serum urate in these results is perhaps unsurprising, given that the Olink panels are populated with previously recognised disease biomarkers. However, these analyses have taken a high-dimensional dataset and reduced it to a dozen proteins with distinct roles, all of which appear to be independently associated with serum urate. The majority of these are positive associations, which is to be expected given the pro-inflammatory nature of serum urate and its elevation in CVD.

The range of accepted roles played by these phenotypes is diverse, including phosphate metabolism and bone development, glucose metabolism, adipocyte function, blood pressure regulation and in the case of the lipid associations, possibly cognitive function. In some cases, such as FGF-23, these results reinforce previously reported findings, showing that they persist even after accounting for several hundred covariates. In others, such as FABP4 and PON3, they explicitly identify associations that were hinted at in work in on related proteins or in different species. And in many, there is no previous report of a link to serum urate. Future confirmation of these links may reveal novel aspects of the biology of serum urate.

A key question that this set of phenotypes raises is why they specifically are associated with serum urate levels, when the dataset includes many other proteins of similar function. With the data currently available, it is difficult to answer this. Interestingly, a recent publication investigating associations between proteins on Olink CVD-I, CVD-II and CVD-III and lipid fractions identified significant associations with six of the eleven urate-associated proteins I have found – CCL3, FABP4, FGF-23 and IL-1RA are all positively associated with triglycerides (TG) and negatively associated with HDL-cholesterol, CHI3L1 positively associated with TG and negatively with LDL-cholesterol, and REN is positively associated with TG only. None

of these proteins were found to be causally associated with lipid fractions in MR analysis[201]. Of course, the nature of the Olink panel is that it contains a relatively small set of proteins known to be related to cardiovascular disease or inflammation, so the set of possible proteins which could be found is already enriched for 'interesting' combinations. Nonetheless it is intriguing to see such a large proportion of the same proteins appearing. Uric acid is positively and strongly correlated with TG. An MR study in the ARIC and FHS cohorts found no evidence that serum urate levels were causal[202], although the authors suggested that dietary sugar intake could be a confounding factor.

When the SCALLOP consortium completes and publishes its ongoing GWAS of the INF, CVD-II and CVD-III panels it should become possible to perform mendelian randomisation analyses with serum urate – this may provide evidence of causality, which should guide further experiments. If any of these phenotypes prove to be a consequence of serum urate levels, it may mean that urate-lowering medication could be an option for treatment in the diseases with which they are associated.

Chapter 3 GWAS of serum urate

3.1 Background

3.1.1 Genome-wide association studies

Genome-wide association studies, commonly known as GWAS, have been referred to elsewhere in this thesis already, but it will be briefly expanded upon here. The purpose of a GWAS is to identify variation in the genome which significantly influences a phenotype. The central precept is that by identifying these genomic regions, we can identify biological mechanisms affecting the phenotype, information that can lead to a better understanding of health and disease. Phenotypes can be binary (e.g. disease status) or quantitative (e.g. concentration of serum uric acid).

Most GWAS are run on single-nucleotide polymorphisms (SNPs), single-base locations in the genome that vary between individuals. Each SNP has two or more possible alleles, the frequencies of which are not necessarily equal and often vary between populations. Individuals can be homozygous at a SNP (with the same allele on each chromosome) or heterozygous (with different alleles on each chromosome). The exception to this rule is the non-pseudoautosomal region of the X chromosome in men, where only one copy of each SNP is possible. Generally, the contribution of each SNP across the genome is assumed independent of all other SNPs, allowing their effect on the trait to be assessed one at a time, greatly simplifying the statistical analysis.

If one SNP allele is significantly more common in cases vs controls, or in individuals with high values of a quantitate trait vs low values, that SNP can be said to be associated with that trait. For quantitative traits, a generalised linear model (GLM) is used to carry out the association test. Commonly and in unstructured samples, the trait value is simply regressed on the number of the chosen reference allele at a SNP (0, 1 or 2). Analysis of Variance (ANOVA) test which is equivalent to a linear regression is also being used frequently to test differences in mean trait values across genotype categories at the SNP (without any assumption on the relationship between these means). For binary traits in unstructured samples, a logistic regression is often used. GLMs allow adjustment for covariates known to be associated with variation in the trait – age and sex are almost ubiquitously adjusted for, as most biological traits vary with both. More complicated adjustments include the use of mixed models to

account for population structure – related individuals within the population, who may have shared environments and genotypes that could confound the association between genotype and the trait of interest. The generally accepted threshold for a significant association after testing the whole genome is $P < 5 \times 10^{-8}$ – this is approximately equivalent to a Bonferroni-adjusted threshold of $P < 0.05$, based on a million independent tests being made. This is not equal to the number of SNPs being tested, as many SNPs are in LD with one another, meaning many tests are not independent.

In the simplest case, one allele at a SNP in a coding region of the genome may directly alter the structure of a protein that regulates the phenotype of interest and change its function. Associations of this sort are generally strong, easier to detect, and above all rare in GWAS results. Most variants associated with a common trait/disease are expected to locate in non-coding regions where the causal variant(s) may alter gene regulation by modifying a binding site for a transcription factor or affecting local chromatin structure. Associated variants can also be in regions of the genome far from any coding genes, making it difficult to interpret the signal.

A further complication due to the structure of the human genome is that the causal variant is in LD with a number of other variants – meaning individuals tend to keep the same combinations of alleles across several SNPs in a given region due to recombination hot spots distribution, creating correlations of allele across loci. This means that if one causal SNP is associated with the phenotype, several SNPs across an LD block are as well. Thus, the characteristic signature of a GWAS 'peak' a region of SNPs clustered together, all associated with the trait of interest. These peaks give 'Manhattan' plots their name – when negative $\log_{10}$-transformed P-values for SNPs are plotted in order of their location along the genome, the resulting plot will tend to have several tall stacks of points that loosely resemble the Manhattan skyline (if the GWAS is sufficiently powered to detect associations). This set of SNPs may contain a causal variant, or several such variants, but sometimes the 'true' causal SNP is not genotyped, though its presence may be inferred from the significance of the SNPs in LD with it. Narrowing down the set to identify the most likely causal variants is known as 'fine mapping'.

A GWAS's power to detect an association between a variant and a phenotype depends on four things – the LD between the assessed variant and the causal SNP, the size of the effect, the frequency of the allele tested and the size of the sample.

Effect size and allele rarity are essentially fixed (within a certain population, though both may vary between populations), so the easiest way to increase the power of a GWAS is to recruit more individuals.

The second way to increase the ability of a GWAS to detect associations is to increase the number of markers tested across the genome. This increases the chances that a marker in high LD with the true causal SNP will be tested (or that the true casual SNP will be included). Many studies are turning to whole-genome sequencing to obtain denser information, but this remains expensive, limiting sample size, though it does have the advantage of allowing *de novo* mutations to be included in the GWAS. Most studies have historically used comparatively cheap genotyping 'chip' arrays that give the genotype at a pre-defined set of several hundred thousand SNPs. One could obtain denser marker information by re-typing samples on newer, denser chips, but this is expensive in terms of both money and biological samples. Instead, it is possible to use information about LD structure in the population of interest to statistically infer the genotype at a large number of untyped markers, a process known as genotype imputation. With a sufficiently dense reference panel, it is possible to impute a population genotyped with a few hundred thousand SNPs to as many as 40 million variants with a high degree of accuracy. This has the added benefit of allowing studies genotyped on a range of different arrays to be imputed to the same set of SNPs for meta-analysis. Almost all current GWAS studies are run using imputed genotype data.

3.1.2 GWAS of serum urate

Over the last decade GWAS have become a major component of statistical genetics. Uric acid has been studied in GWAS since the field's inception. The detection of a strong association at the *SLC2A9* locus[66] and subsequent discovery that its encoded protein GLUT9 was in fact a urate transporter[55,56] was an early example of the utility of the technique. GWAS have been performed on serum urate since 2007, advancing from single cohort studies genotyped on custom SNP arrays to consortium meta-analyses using imputation panels to boost their power.

A landmark meta-GWAS of serum urate was conducted on a discovery sample of 110,347 European-ancestry individuals and replicated in a sample of 32,813 by the Global Urate Genetics Consortium (GUGC), published in 2013[74]. Genotypes were imputed to version 2 of the HapMap reference panel[203], and results were reported for around 2.5 million SNPs. This analysis identified 28 loci, of which 18 were novel. Since

then, there have been two major new imputation reference panels, first using the 1000 Genomes Project[147], and the second from the Haplotype Reference Consortium. The HRC panel is much denser than the HapMap2 panel – over 39 million SNPs after quality control filtering – allowing finer mapping of associated variants, and allows accurate imputation of variants as rare as 0.1% minor allele frequency (MAF)[204].

The most recent large GWAS of uric acid, at the time of writing, is in the Biobank Japan cohort. Serum urate was one of 58 quantitative traits regressed against 5.7 million SNPs imputed to the 1000 Genomes Phase 1 version 3 panel in 109,029 individuals of Japanese ancestry. The analysis identified 27 loci, of which 10 were novel[77]. This demonstrates that GWAS of comparable size using samples from different ethnic groups can detect quite different sets of signals, although in this case the use of a denser imputation panel makes it difficult to distinguish which new hits are due to ancestry.

It has been shown that even on the same sample a newer, denser imputation panel can identify signals that are missed when compared to a less dense panel[205]. This can be through imputation of variants in regions previously missed, through closer tagging of causal variants or through more accurate imputation of SNPs. A new large-scale serum urate consortium analysis has been orchestrated by the CKDGen consortium to take advantage of both the newer imputation panels and increasingly large number of cohorts willing to contribute data. My contribution to this project will be the subject of Chapter 4.

However, large consortium analyses are slow to progress, due to the administrative and analytic burden of working with so many cohorts. I have access to uric acid data from several in-house population cohorts, including the CROATIA-Vis, CROATIA-Korcula, ORCADES, VIKING and Generation Scotland: Scottish Family Health Study[206] cohorts – combined this represents a sample of over 10,000 individuals, more than sufficient to identify new associations from the HRC imputation panel. Furthermore, all but Generation Scotland are isolate population cohorts – gathered from isolated populations where genetic drift can increase the frequency of rare variants to detectable levels. While the focus of the CKDGen meta-analysis is on identifying transethnic variants, SNPs which have detectable effects in individuals from any genetic background, a smaller meta-analysis enriched for isolate populations stands a chance of detecting signals driven by low frequency variants which may be overlooked in the larger analysis.

The aim of this chapter is to explore the potential of a meta-analysis of our own cohorts, imputed to the HRC reference panel, and establish whether I can detect any novel associations that the CKDGen consortium might miss.

3.2 Methods

3.2.1 Cohorts

3.2.1.1 CROATIA

In addition to the CROATIA-Vis cohort described in Section 2.1.1.1, there are two other separate studies that are part of the 10,001 Dalmatians project[122]. CROATIA-Korcula (Korcula) is another isolate population study, similar to Vis, but conducted on the island of Korčula. A total of 2,832 participants were recruited in three phases between 2007 and 2014. Uric acid measurements from collected serum aliquots are available for 2,673 participants.

In both Vis and Korcula, serum urate was measured using the uricase UV photometry method in the Labor Centar Biochemical Laboratory, Zagreb Croatia.

3.2.1.2 ORCADES

The ORCADES cohort is described in Section 2.1.1.2. 1,964 participants had both serum urate measurements and genotype data. Serum urate was measured using the uricase/peroxidase method in the Balfour Hospital, Kirkwall, UK.

3.2.1.3 VIKING

The Viking Health Study (VIKING) is another Scottish isolate population, from the Shetland isles. Similar to ORCADES, though slightly less stringent, participants were recruited if they had at least two grandparents from the Shetland isles. A total of 2,105 participants were recruited between 2013 and 2015. Measurements are available for 2,090 individuals. Serum urate was measured using the uricase/peroxidase method in the Gilbert Bain Laboratory, Lerwick, Shetland.

3.2.1.4 Generation Scotland

The Generation Scotland: Scottish Family Health Study (Generation Scotland, or GS) is the largest cohort within the Generation Scotland project. It is a general population cohort with enrichment for families – recruitment was both through GPs and by directly contracting first degree relatives of people who had already been recruited. The cohort comprises 24,000 individuals, of which 20,032 have genotype data available[206].

Uric acid was not included in the biochemical measures performed in this cohort. However, permission was obtained from 23,603 participants to link their GS data to their electronic health record (EHR) data held by the Scottish National Health Service (NHS Scotland), using their Community Health Index (CHI) number, detailed below.

3.2.1.4.1 Generation Scotland Electronic Health Record Linkage

Record linkage was performed by Dr Shona Kerr and Archie Campbell (University of Edinburgh), and I processed and analysed the data once records had been linked to the genetic data held by Generation Scotland.

EHR linkage provided access to several databases of information, including the results of blood biochemistry tests. EHRs were first obtained from record linkage to 11,125 Generation Scotland participants in the NHS Tayside region, where 21% of participants have had at least one test for uric acid. Subsequently, access was obtained for 8,264 individuals the NHS Greater Glasgow and Clyde region with 2,559 separate measurements for uric acid. Data from the final region, NHS Aberdeen, was not available at the time this analysis was conducted. All serum urate measurements were performed in NHS laboratories according to NHS standard procedures, but no information is available on the specific assays used.

In GS Tayside, 6,268 measures of uric acid were available for 2,356 individuals. Where multiple tests were available, the highest measurement was selected for analysis on the basis that this likely reflected the level prior to any intervention. Pregnant individuals were removed. Of these, 2,077 individuals were genotyped and also passed the necessary filtering to be included in the HRC-imputed data set (as described in Nagy *et al.*[206]) and were used in the analysis. Age at measurement was calculated using reported date of birth and date of blood test. After applying the same criteria to measurements from GS Glasgow, 1,155 individuals were included in the GWAS.

The results from the GS Tayside analysis were included in a descriptive paper of the cohort that was published separately[206]. These were analysed separately and compared to the results of a GWAS run only on the genotyped SNPs, to explore the potential of HRC imputation. The genotyped SNP GWAS was run entirely within the R *GenABEL* suite, as the number of SNPs was small enough to make this computationally efficient, while the imputed GWAS was run as described in Section 3.2.3.

3.2.2 Genotype imputation

Genotyped SNPs were imputed to the HRC.r1-1 reference panel by Thibaud Boutin, using ShapeIt2 for data phasing and the Sanger Imputation Service server[204] for imputation. Monogenic variants and variants with imputation quality of less than 0.4 were filtered out, leaving 24,111,857 variants for analysis.

3.2.3 Regression analysis

Analyses were run on each cohort individually, with the two Generation Scotland subsets treated as separate cohorts – this was a practical decision made because of the multi-stage data release. Three analyses were run per cohort – females-only, males-only and sex-combined.

Phenotypes were adjusted for age and sex (in the sex-combined analysis only) and corrected for relatedness. This was performed with the GRAMMAR-Gamma method[207], implemented in the polygenic function in *GenABEL*.

GRAMMAR-Gamma is a computationally-efficient two-step method where phenotypes are adjusted for covariates and a genomic kinship matrix is fitted. The residuals from this are used as outcomes in the actual regression analysis. A correction factor is also calculated. In the second step, these residuals are regressed on genotypes, and the regression coefficients and P-values obtained are corrected using the GRAMMAR-Gamma correction factor, which adjusts for the biased effect estimates produced by the GRAMMAR models. The regression of the genotypes on the residuals was performed using RegScan v0.2[208].

3.2.3.1 Meta-analysis

Meta-analysis was conducted in METAL (v2011-03-25)[209] using the standard-error-weighting method to combine effect sizes. Post-analysis results were filtered to retain only those SNPs with a Minor Allele Count (MAC) > 20 per cohort and which were present in at least two cohorts. The standard GWAS threshold of p-value ≤ 5×10^{-8} was used to identify significant associations. Loci were defined as the region of ±500kb around the SNP with the lowest P-value in the region – the 'index SNP' for that locus. Index SNPs with overlapping loci were merged and the SNP with the lowest P-value retained as the index SNP. Loci are named for the nearest gene to the index SNP, obtained from the Ensembl Variant Effect Predictor (VEP)[40]. Each locus was checked against the summary statistics from the 2013 GUGC meta-analysis from

Köttgen *et al.*[74]. Any locus which contained one or more significant SNPs in the GUGC meta-analysis was considered to have been identified in that analysis, although the name of the region was not always the same.

Heterogeneity across analyses was assessed using the I^2 statistic calculated by METAL, which represents the percentage of variation across studies that is due to heterogeneity rather than chance.

3.2.4 Conditional analysis

To investigate an unusual signal on chromosome 11 (see Results), a conditional analysis was run on the SNP with the lowest P-value in the region, to check whether there were multiple independent signals. This was performed as above, with the only change being that the genotype of the top SNP was included as an additional covariate. Because this analysis was performed before the full meta-analysis was finalised, it did not include the GS Glasgow data, as it was not available at the time, and additionally used a reduced dataset of approximately 950 individuals from Korcula.

3.3 Results

3.3.1 Cohort summary information

Summary statistics for the cohorts included in the meta-analysis are given in **Table 15**. All cohorts are similar in terms of being slightly biased towards females, and the mean serum urate levels are comparable. Standard deviations are slightly higher in the EHR-derived measurements in GS, which could be driven by the selection of the highest measure when multiple values were available for one person. It could also be due to more people being on urate-lowering medication.

Table 15 – Per-cohort summary statistics

			Age		Serum urate (mg/dL)	
Cohort	**N**	**Female (%)**	**Mean**	**SD**	**Mean**	**SD**
Vis	948	58%	56.17	15.52	5.22	1.59
Korcula	2,683	64%	53.96	15.85	4.94	1.35
ORCADES	2,073	61%	53.22	15.29	4.91	1.19
VIKING	2,093	60%	49.89	15.21	5.10	1.20
GS Tayside	2,077	61%	50.59	15.09	5.76	1.85
GS Glasgow	1,155	58%	54.73	14.62	5.75	1.89

3.3.2 GS Electronic Health Records

3.3.2.1 Distribution of phenotype

The distribution of uric acid measurements in the GS EHR cohorts is compared to the combined cohorts measured at baseline in **Figure 22**. While the mean and standard deviation for these cohorts are higher than the others, and the distribution has a longer upper tail, this is to be expected, as the highest measure was always selected for any individual with more than one measurement available. Additionally, because the measurements are obtained from NHS blood tests, the subset of individuals with serum urate measured may be enriched for people with a medical problem. Rank transformation to normality would eliminate this but could also remove genuine biological associations enriched in the GS sub-cohorts that are driving high serum urate levels, so I decided to analyse the data untransformed. The means remain within the normal range of serum urate levels in healthy humans.

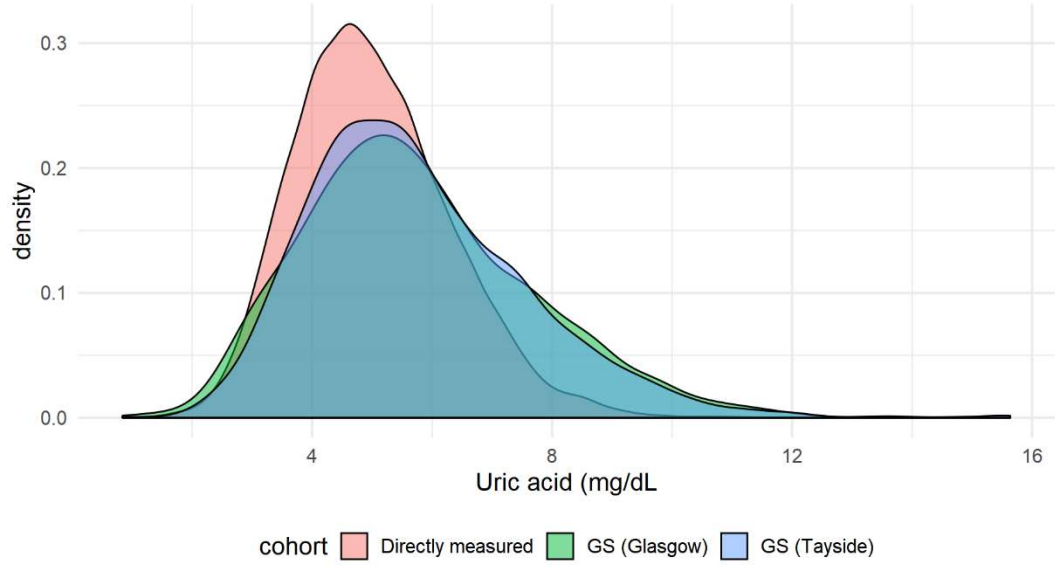


Figure 22 - Distribution of serum urate measurements in GS EHR records compared to all cohorts with directly-measured serum urate.

3.3.2.2 GS Tayside EHR GWAS

The results of the GS Tayside EHR GWAS are shown in **Figure 23**. The strongest peak is in the *SLC2A9* region, which encodes the well-known urate transporter GLUT9. This association is the strongest known GWAS hit for serum urate and serves as a good positive control for the EHR-derived serum urate phenotype. Additional genome-wide significant signals were detected at three additional loci, none of which have been previously reported. The index variants (the SNPs with the lowest P-values within a ±500kb window) for these regions are shown in **Table 16**.

All three novel index SNPs are rare (MAF < 1%) and are present only in the imputed GWAS as they were not directly genotyped. They all have large effects compared to the *SLC2A9* lead SNP but also much larger standard errors due to their rarity (at the limit of our cut-off of MAC > 20).

These results were published in Nagy *et al.* 2017[206] along with several other biochemical and anthropometric traits. The purpose of this paper was to showcase the cohort and the value of the new, denser imputation possible with the HRC panel. The serum urate results additionally highlight that the value of the EHR linkage for normal quantitative trait GWAS, showing well that measurements taken for clinical purposes, replicate results from more traditionally measured cohorts.

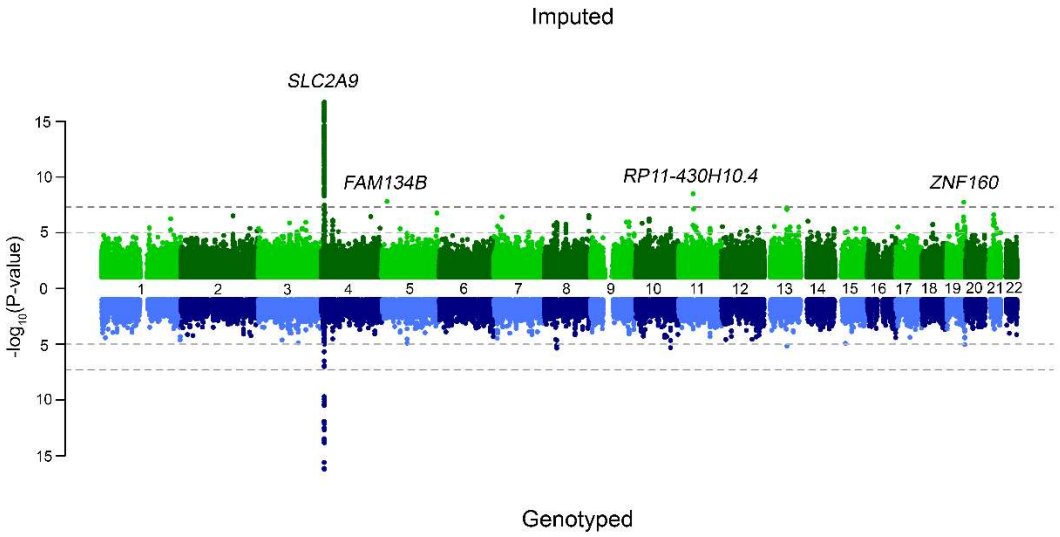


Figure 23 - Miami plot of GS Tayside uric acid results

The top panel shows the results using SNPs imputed to the HRC reference panel, while the bottom panel shows only directly genotyped SNPs. Genome-wide significance is indicated by the dark grey dashed line and the suggestive threshold by the light grey. Labels are the closest gene to the top SNP. SNPs with P-value >0.01 are not shown for clarity. (This figure is published in Nagy et al. 2017[206].)

Table 16 – Genome-wide significant index SNPs from the GS Tayside serum urate GWAS.

SNP	MAF	Effect	Std. Err.	P-value	Gene	Imputation Quality
rs6449213	0.1652	0.592	0.070	1.93×10^{-17}	*SLC2A9*	1.00
rs75869162	0.0054	2.245	0.397	1.57×10^{-08}	*FAM134B*	0.80
rs141208451	0.0053	2.319	0.391	3.13×10^{-09}	*RP11-430H10.4*	0.86
rs187171029	0.0060	1.999	0.355	1.84×10^{-08}	*ZNF160*	0.91

3.3.3 Meta-analysis

The sex-combined meta-analysis included a total of 10,908 individuals, and 6,622 and 4286 in the female and male analysis respectively. The results of these are plotted in **Figure 24**. The index SNPs and details of the loci are reported in **Table 17**.

The strongest association was, as expected, with *SLC2A9. ABCG2* was detected in sex-combined and males but not in females, despite the larger sample size. This is in agreement with the known stronger effect in males than in females at this locus.

Six loci were identified in the sex-combined analysis, of which two were not reported in the GUGC 2013 paper despite the sample size being considerably smaller in this analysis. Four were identified in the female-specific analysis, three of which were not in the GUGC results, and four in the male-specific analysis, of which one was not in the GUGC results.

All the new loci are driven by rare index SNPs (MAF < 1%) with the exception of rs139498948 at *SYT1* in males, which is still uncommon at MAF = 1.1%. None of these SNPs are present in all six cohorts, and many of the loci show high heterogeneity (I^2 > 50%). *OR52I1* and *CDC42BPB* in females and *SYT1* in males are below this threshold.

Forest plots for the index SNPs at these loci are shown in **Figure 25**, as well as the index SNP for *FCHSD2*, which has an I^2 value of 51.6%. These show that the signal in at rs573624409 in the *FCHSD2* locus seen in the in sex-combined analysis is only significant in the ORCADES cohort, though the direction of effect agrees across all cohorts. The signal at rs1876025 in *OR25I1* in females is significant in Korcula and VIKING but may be partially driven by an implausibly strong effect in GS Glasgow with a very large standard error. Also in females, rs576082236 in *CDC42BPB* shows a consistent, significant signal between VIKING and Vis, but non-significant effects in the opposite direction in the two GS cohorts. In males, rs139498948 at *SYT1* has quite high heterogeneity, but the direction of effect appears to be consistent across studies.

None of the index variants in the novel female-specific loci were significant even at a nominal level (P < 0.05) in males, but the index variant in *SYT1* was nominally significant in females (Effect: 0.466, Std. Err:0.165, P = 0.00472).

None of the three index variants from the GS Tayside analysis reach genome-wide significance in the meta-analysis, but two are significant at a lookup level (p < 0. 05 / 3) and the third is nominally significant (**Table 18**). Effect directions match in all three cases, although the size of the effect is considerably smaller. However, in all three cases, heterogeneity is very high (I^2 > 85%) suggesting these loci may not be reliable. Forest plots of the effect sizes from each cohort (**Figure 26**) confirm that the signals in the meta-analysis appear to be driven in large part by GS Tayside. In the case of rs141208451, ORCADES has a similar effect, but this falls in the unusual region on Chromosome 11 described below.

A wide area of significant SNPs with overlapping loci was identified in on Chromosome 11. This encompassed, amongst others, the known urate loci *SLC22A11*, *SLC22A12* and *NRXN2*. To understand this unusual pattern, results for this chromosome were plotted separately for each cohort in **Figure 27**. The strong signals seen in the meta-analysis appear to be being driven by a wide region of associations in the ORCADES cohort. This region is unusual in that it not only spans almost 10Mb but also crosses the centromere. There is also a signal in the GS Tayside sample, this is the *RP11-430H10.4* locus reported in Section 3.3.2.2.

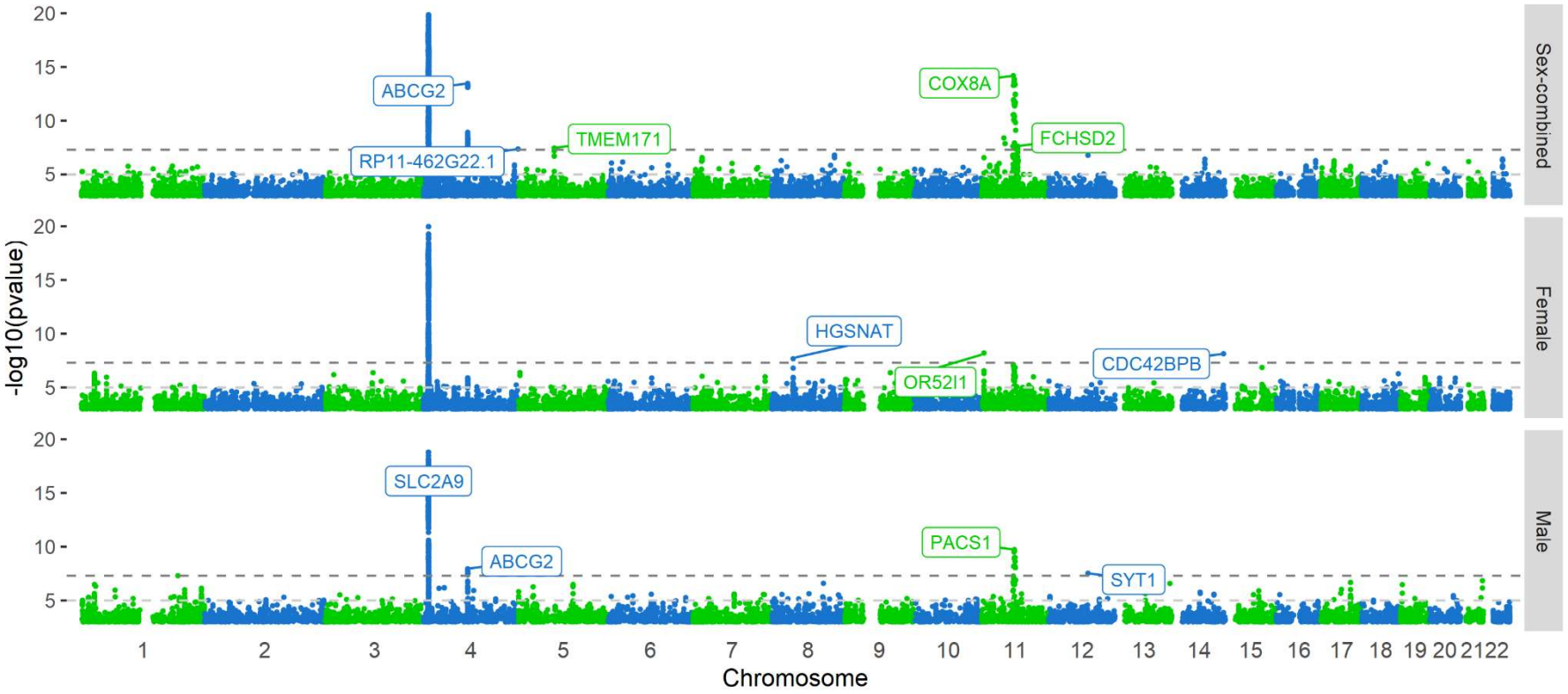


Figure 24 - Manhattan plots of uric acid meta-GWAS.

Labels are index SNPs, defined as the SNP with the lowest p-value in a window of ±500kb. Labels are the gene symbol of the closest gene, identified with the Ensembl VEP. Y-axes are truncated at 20 for clarity – *SLC2A9* index SNP is consequently not shown on the Sex-Combined and Female analyses. Dark grey dashed line is the genome-wide significance threshold, light grey dashed line is the suggestive significant threshold.

Table 17 – Index SNPs from the sex-combined and sex-separate meta-analyses.

In All, the *COX8A* locus contains 9 index SNPs with overlapping regions from 11:63,032,537 – 11:68,562,669, including one mapped to *SLC22A12*. In Males, the *PACS1* locus contains 6 index SNPs with overlapping regions from 11:64,359,252 to 11: 67,846,680, also including one mapped to *SLC22A12*. Loci not reported in Köttgen *et al.* 2013 are highlighted in gold. Frequency of A1 is given for the meta-analysis population.

	Index SNP rsID	Chr.	Position	Locus	SNPs in locus	A1	A2	N	Freq A1	MAF	Effect	Std. Err.	P-value	Cohorts	I^2 (%)	In GUGC 2012
All	rs3775947	4	9,995,240	*SLC2A9*	1,194	T	C	10,907	0.777	0.223	0.376	0.020	7.20E-81	6	19.80	Yes
	rs1481012	4	89,039,082	*ABCG2*	46	A	G	10,907	0.897	0.103	-0.211	0.028	3.39E-14	6	54.60	Yes
	rs577353818	4	190,587,150	*RP11-462G22.1*	1	T	C	5,322	0.998	0.002	-2.488	0.454	4.32E-08	3	85.80	No
	rs13182742	5	72,425,458	*TMEM171*	1	T	C	10,907	0.746	0.254	0.111	0.020	3.46E-08	6	0.00	Yes
	rs542534688	11	63,747,858	*COX8A*	27	T	C	8,234	0.008	0.008	-1.031	0.132	6.53E-15	5	0.00	Yes
	rs573624409	11	72,799,547	*FCHSD2*	2	A	G	7,286	0.994	0.006	0.895	0.160	2.34E-08	4	51.60	No
Female	rs9994216	4	9,984,541	*SLC2A9*	1,120	T	G	6,622	0.778	0.222	0.408	0.024	1.22E-63	6	52.90	Yes
	rs112029032	8	43,054,647	*HGSNAT*	1	A	G	3,735	0.005	0.005	1.830	0.327	2.14E-08	4	72.00	No
	rs1876025	11	4,615,987	*OR52I1*	1	A	C	4,886	0.998	0.002	-1.758	0.303	6.33E-09	4	0.00	No
	rs576082236	14	103,468,041	*CDC42BPB*	1	A	G	3,735	0.003	0.003	3.938	0.681	7.35E-09	4	11.40	No
Male	rs3775947	4	9,995,240	*SLC2A9*	179	T	C	4,286	0.774	0.226	0.308	0.034	1.57E-19	6	0.00	Yes
	rs3109823	4	89,064,602	*ABCG2*	10	T	C	4,286	0.720	0.280	0.182	0.032	1.13E-08	6	0.00	Yes
	rs185509475	11	65,943,116	*PACS1*	14	T	C	2,430	0.010	0.010	-1.827	0.287	1.87E-10	3	0.00	Yes
	rs139498948	12	79,289,619	*SYT1*	1	A	G	2,911	0.011	0.011	1.184	0.214	2.97E-08	4	47.60	No

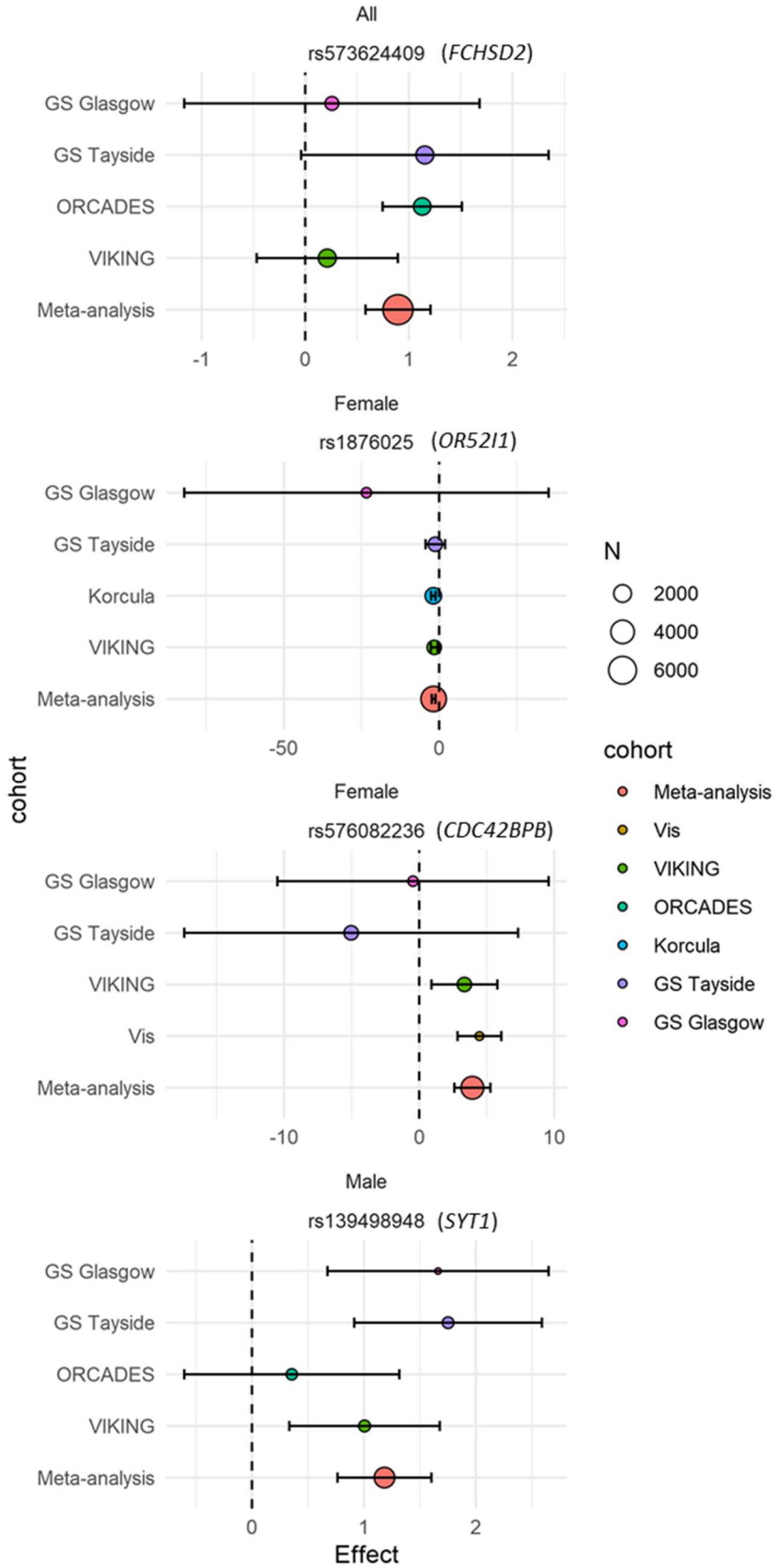


Figure 25 - Forest plots of novel index SNPs with I2 < 60%.

Table 18 – Lookup of GS Tayside novel loci in meta-analysis results.

rsID	Chr.	Pos.	Locus	EA/NEA	N	Freq. A1	Effect	Std. Err.	P-value	I2 (%)
rs75869162	5	16,617,922	*FAM134B*	A/G	7,286	0.004	0.419	0.181	0.02079	89.4
rs141208451	11	45,538,920	*RP11-430H10.4*	A/G	7,286	0.008	0.429	0.150	0.00434	91.3
rs187171029	19	53,599,256	*ZNF160*	T/A	7,286	0.007	0.488	0.140	0.00048	87.3

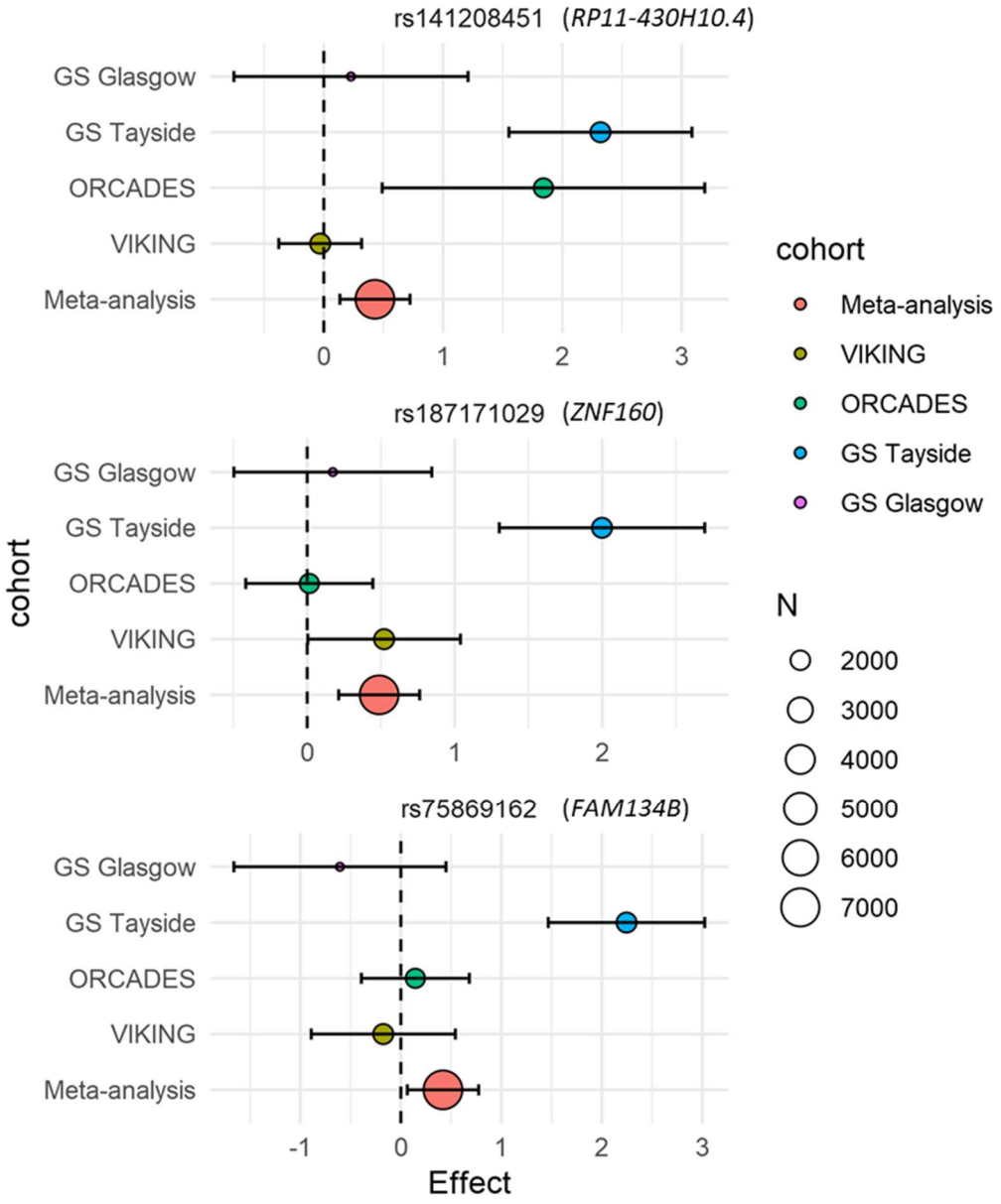


Figure 26 - Forest plot of novel index SNPs from GS Tayside GWAS.

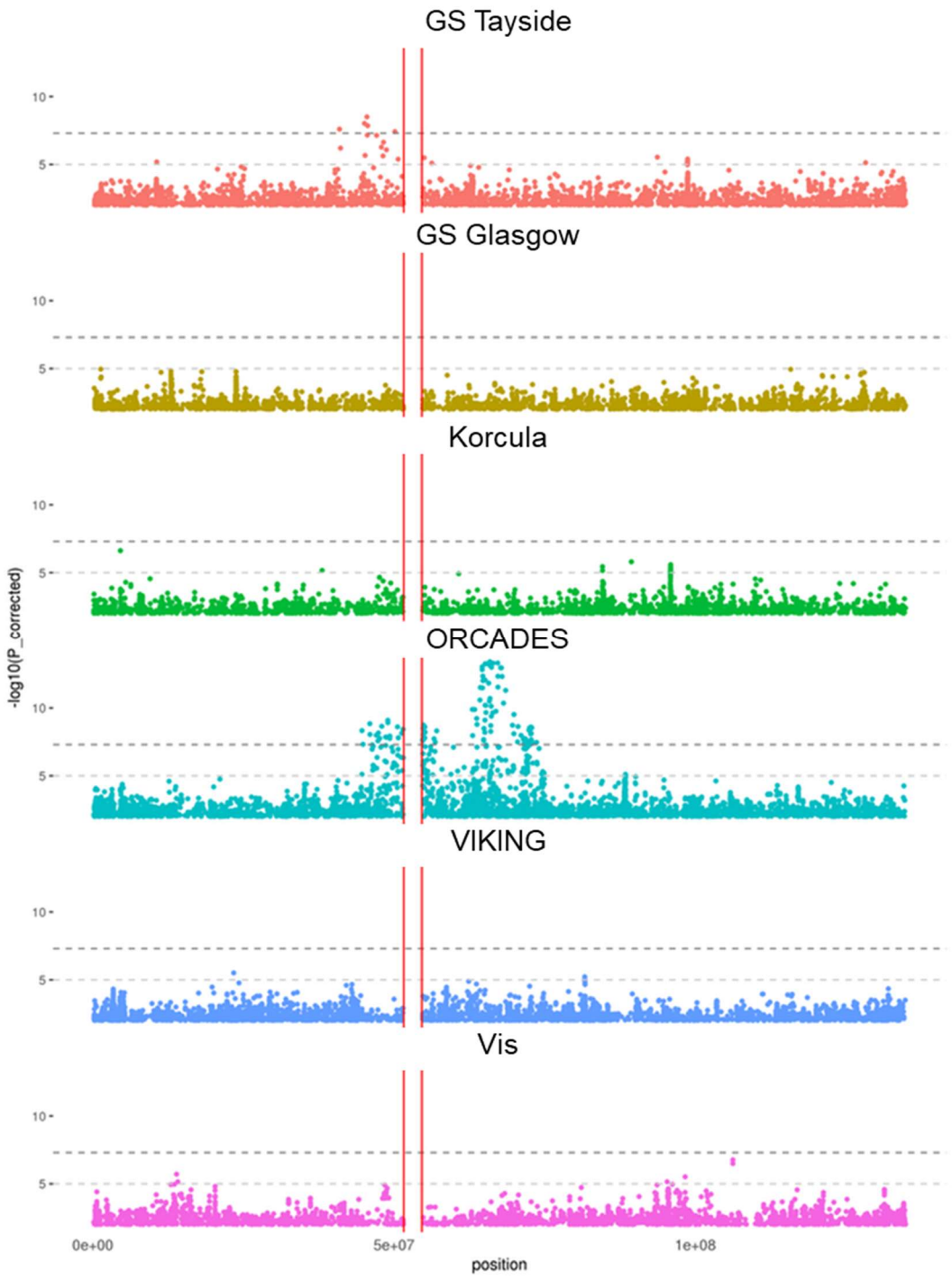


Figure 27 – Per-cohort Chromosome 11 Manhattan plots.

Only SNPs with P < 0.001 and MAC > 10 are shown. Dark grey dashed line is the genome-wide significance threshold, light grey dashed line is the suggestive significant threshold. The centromere is delimited by the red lines.

3.3.4 Conditional analysis

The SNP with the lowest p-value on Chromosome 11 was rs542534688 in the *COX8A* locus (**Table 17**). Conditional analysis on this SNP was run in the ORCADES cohort, which removed all significant signals on Chromosome 11 (**Figure 28A**).

Conditional analyses on rs542534688 were also run in Vis, ORCADES and GS-Tayside cohorts along with a subset of the Korcula cohort. These were meta-analysed for a total of 8,164 individuals, which also showed no signals on Chromosome 11 (**Figure 28B)**. This suggests that broad signal identified in the meta-analysis is tagging a single causal region, or multiple regions in high LD.

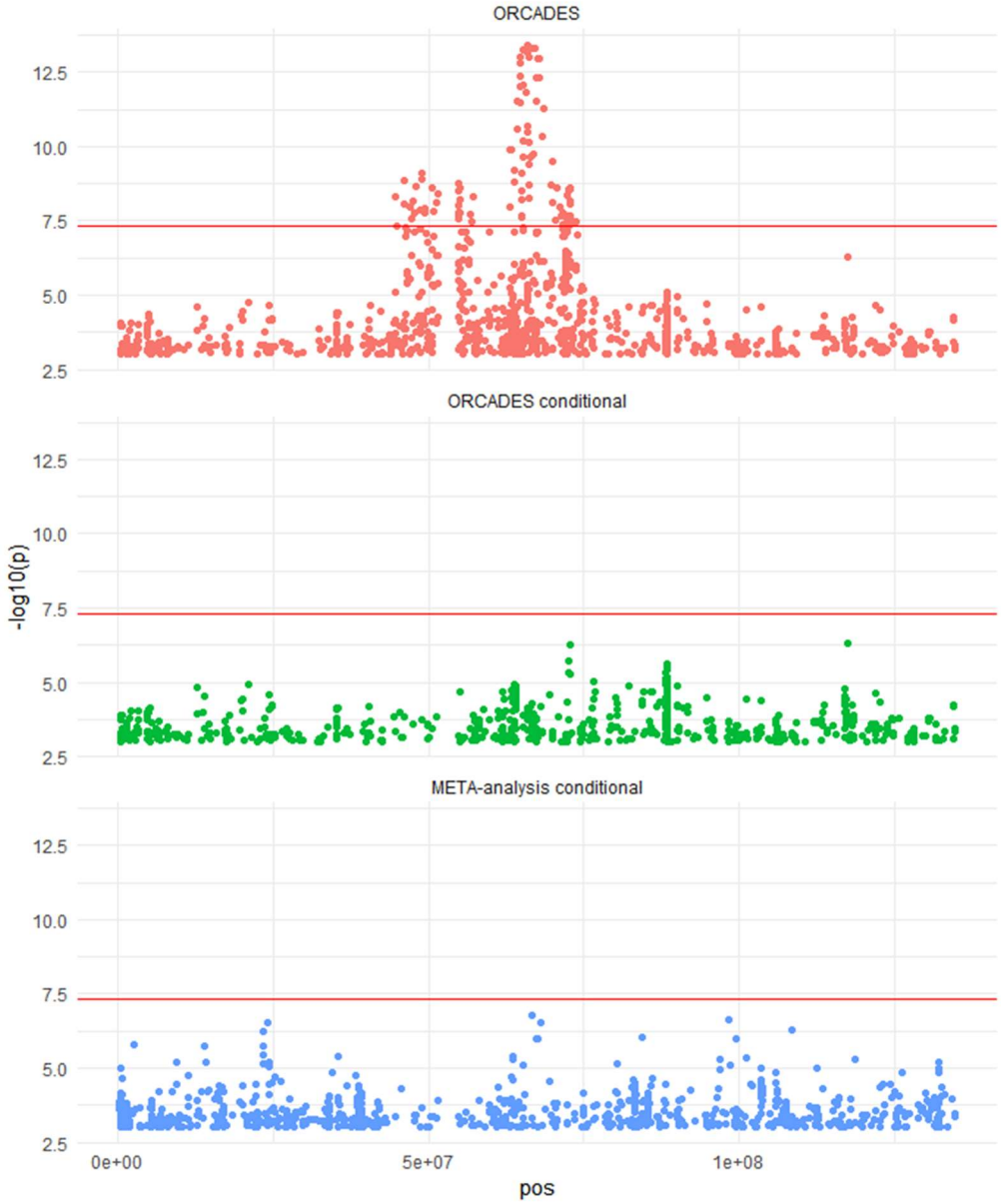


Figure 28 – Manhattan plots for chromosome 11 of conditional analysis on rs542534688.

Top panel: in the ORCADES cohort alone with no conditional analysis; middle panel: in ORCADES conditioned on rs542534688; bottom panel: in a conditional meta-analysis of 8,164 individuals. Only SNPs with P-value < 0.001 and MAF > 0.001 are shown. Red line denotes the threshold for genome-wide significance ($P = 5x10^{-8}$).

3.4 Conclusions

3.4.1 GS Tayside EHR linkage

The results of the GS Tayside GWAS demonstrate the value of health record linage for biobanks. For the minimal cost of an access fee, and a small amount of data processing time, we have obtained an additional phenotype for a fraction of the cost of more traditional recall of participants. The results of the GWAS replicate the expected hits for a cohort of this size, addressing possible concerns that the subset of the population with blood biochemistry data available will have abnormal measurements due to illness. I also identified loci which had not been previously reported, though these loci showed high heterogeneity in the meta-analysis and may be driven largely by the GS Tayside sub-cohort alone.

Serum uric acid is only one of a large number of phenotypes that are available through the clinical test results, which include several which are not available in many population cohorts. There is also the potential for time-series data, which was not explored here. Many individuals have multiple blood tests over time, which if sample sizes were sufficiently large and suitable controls could be identified, could be used to assess the effects of medication or treatment. The UK Biobank, an even larger cohort than GS at over 500,000 individuals, has been designed with record linkage in mind – though uric acid is unfortunately not available (as yet) either directly or through health records. Over time, as the participants age, it should yield sufficiently large numbers of individuals with health issues to investigate most diseases that are common in the UK without the need to design specific cohorts – and crucially, with information about individuals before they developed their disease. EHR studies are unlikely to – and should not – replace population cohorts entirely, which assess phenotypes in a standardised and high-quality manner, and include healthy individuals rather than those who have sought medical care. Rather they will provide an increasingly valuable source of additional information on the natural history of diseases.

3.4.2 Serum urate meta-analysis

This meta-analysis demonstrates the value of a denser and more accurate imputation panel for discovery of variants associated with serum urate levels. The most recently-published large-scale meta-analysis, Köttgen *et al.* (2013), was performed on version 2 of the HapMap imputation panel[203], which covers approximately 2.5 million SNPs,

most of which have a MAF over 5%. By contrast, the meta-analysis performed here on genotypes imputed to the HRC.r1-1 panel included nearly 22 million SNPs after filtering for imputation quality. The imputation accuracy of rare SNPs is much higher, allowing the use of much lower filters for minor allele frequency – here a cut-off of MAF > 0.001 was used, corresponding to approximately 20 copies of the minor allele across over 10,000 individuals. The improved performance of the HRC imputation panel has been documented before, but at the time of analysis, this meta-analysis was the first time that uric acid had been regressed on HRC-imputed genotypes.

The strongest signals are unsurprisingly from the known and well-characterised serum urate loci *SLC2A9* and *ABCG2*, as well as *SLC22A11* and *SLC22A12*. However, the fact that a locus has previously been reported does not mean that its detection is not noteworthy – in particular, the locus at *COX8A* is interesting. This locus was identified in Köttgen 2013 with a sample size of over 140,000 individuals, but the earlier publication by Kolz *et al.* 2009[171] (imputed to the HapMap version 1 imputation panel) did not detect any signal in the region despite a sample size of over 28,000, nearly three times the size of this study – a sample that included the Vis and ORCADES cohorts. Although the signal in my meta-analysis is only just borderline significant, it showcases the value of a denser imputation panel. Even though the sample size was relatively small, the association with *COX8A* was picked up where previous panels have missed it. This may be because the causal SNP linking the locus to serum urate levels is in higher LD with a SNP on the HRC imputation panel – or even present on the panel – leading to a less noisy relationship between genotype and phenotype, which can be detected with a smaller sample.

The *COX8A* locus encompasses the broad signal seen on Chromosome 11, which appears to be driven primarily by the ORCADES cohort. There is some evidence of a signal in this region in GS Tayside at *RP11-430H10.4*, on the other side of the centromere from *SLC2A9*, but this much narrower signal does not look unusual for a GWAS. The pattern in ORCADES could be caused by a pericentromeric inversion in a few related individuals in the cohort, increasing the LD between all SNPs in the inversion. A serum urate-raising variant in *SLC22A11* or *SLC22A12* in these individuals would result in associations with serum urate levels being detected with any SNP associated with the inversion haplotype. Further investigation of this signal will soon be possible once sequencing data is available in the ORCADES cohort. The first step will be to check if the variants are correctly imputed. If the signal appears to

be real, it may be possible to identify structural variants in the region that might explain the unusual signal.

RP11-462G22.1, also known as *LINC01262*, is a long non-coding RNA (lncRNA) on Chromosome 4. The functions of lncRNAs are often not well explored, but transcript levels of this particular lncRNA have been reported to be increased in PD patients, in both RNA-seq analysis in leukocytes and in RT-PCR in tissue from the substantia nigra, one of the relevant brain tissues in PD. Serum urate levels are reduced in the substantia nigra in PD patients[211], and it has been proposed as a biomarker for both risk and progression of the disease[212]. The association detected here is a large serum urate-reducing effect from the major allele, which is common (EAF = 99.8%). It must be noted that this signal is not highly significant, and GWAS hits with borderline significance and low minor allele frequency should always be treated with much caution. The biological connection is compelling, but could easily be coincidence, particularly as uric acid is a prolific phenotype which is linked to a large number of different pathways. Unfortunately, the GTEx database reports no eQTLs for *RP11-462G22.1*, making it difficult to establish whether the SNP identified is also an eQTL for the lncRNA.

FCHSD2 encodes FCH and double SH3 domains protein 2. This gene has been associated with systemic lupus erythematosus[213], an autoimmune disease associated with elevated serum urate levels in patients with normal renal function[214]. The locus has also been associated with inflammatory bowel disease in a Japanese population[215].

Three signals were unique to females. *HGSNAT* encodes Heparan-Alpha-Glucosaminide N-Acetyltransferase, a lysosomal enzyme involved in the degradation of heparin sulphate. Mutations in this gene lead to Sanfilippo syndrome type C, a rare autosomal recessive condition that has no known treatment[216,217]. Symptoms include neurodegeneration, which account for the possible link to uric acid. The gene has also been linked to retinitis pigmentosa[218]. The index SNP, rs112029032, is a missense variant, but is only reported in Phenoscanner to be suggestively associated with “Self-reported bowel or intestinal obstruction” in the UK Biobank[140]. The estimated impact of the mutant allele is inconsistent, with some metrics suggesting it is tolerated and others suggesting it is deleterious. The non-ancestral A allele appears to have a serum urate-increasing effect in women only. Purely speculatively, this could suggest serum urate is increased in females (who have lower serum urate levels) in response

to an increased risk of neurodegeneration, rather than the SNP lowering serum urate levels and leading to neurodegeneration. However, until the deleteriousness of the SNP is established, this is impossible to confirm.

CDC42BPB encodes CDC42 Binding Protein Kinase Beta. This GTPase plays a role in cytoskeletal remodelling in cell growth, and has been shown to phosphorylate non-muscle myosin light chain in rat[219]. I have not been able to find any obvious connection to serum urate, save the somewhat tenuous link that insertions of introns of *CDC42BPB* into other genes has been linked to ankylosing spondylitis[220], a chronic inflammatory arthritis. Low levels of serum urate have been suggested to be linked to risk of low BMD in patients with this disease[221].

OR52I1 encodes Olfactory Receptor Family 52 Subfamily I Member 1. The index SNP is a missense variant, with consensus on Enseml from SIFT, PolyPhen and Mutation Assessor scores that it is likely deleterious to function. It has no associated phenotypes in Phenoscanner. *OR52I1* is under-expressed in substantia nigra in PD patients[222]; as mentioned above, uric acid is also reduced in this tissue in PD.

The only locus uniquely identified in males was *SYT1*, which was also nominally significant in females and consistent in direction of effect. The gene encodes Synaptotagmin 1, a membrane-anchored calcium sensor that plays a key in neurotransmitter release at the synapse[223]. SYT1 is highly expressed in the brain and central nervous system. A GWAS of serum creatinine levels (which included ORCADES and Vis in its discovery sample and Korcula in its replication) identified variants in the *SYT1* gene, and the authors suggested that its expression in renal podocytes suggested it may be involved in regulation of glomerular activity[224]. If this is the case, it could explain its association with serum urate.

The downstream analyses on the results of this meta-analysis are relatively basic, limited to closest-gene annotation of index SNPs defined. This is not uncommon for publications within the GWAS field, and was appropriate for this analysis, the purpose of which was to quickly query our in-house data for new signals present in the HRC-imputed data. However, many cases have shown that the closest gene is not always the gene of interest. The following chapter will detail my contributions to the CKDGen Round IV analyses of uric acid and gout, which will include more sophisticated analyses of a much larger dataset.

Chapter 4 CKDGen Consortium meta-analysis of serum urate

4.1 Background

4.1.1 GWAS consortia

Early GWAS efforts used small samples and could therefore identify only strong genetic associations with their traits, with *SLC2A9* being the quintessential example of success for serum urate. However, as it became apparent that increasingly large sample sizes were successfully returning more associations, of smaller effect magnitudes, in a manner expected under a highly polygenic model of complex traits, these individual small cohorts were no longer analysed on their own.

Collecting larger and larger cohorts would require a long time and large investment, though several such investments were made, such as the UK Biobank and the Million Veteran Program – long term projects which would take years to reach fruition. In the first instance, meta-analysis allowed multiple groups to combine their results without needing to share sensitive personal data. Early collaborations evolved into 'consortia', semi-formal organisations centred around particular groups of traits which any cohort with the correct phenotypes was invited to join.

The general approach of a GWAS consortium analysis is a small group of lead analysts, often comprising a mix of senior researchers who guide the project and early career researchers who do most of the analysis. This small group will identify a research question and design an analysis plan that is distributed to collaborating cohorts. This details the exact analyses to be performed to minimise the heterogeneity in analysis results. 'Study-level' analysts – analysts affiliated with each cohort who can access individual level data – will perform the analyses according to the analysis plan, and results are returned to a central location for quality control (QC) and meta-analysis. The level of involvement for each study is typically quite low, as the bulk of the analysis is typically performed by the meta-analysts.

Today, years into the 'GWAS era', most consortia are mature, with established working practices and frequent in-person meetings. As there is only a finite number of cohorts with any given phenotype, some, such as GIANT and CHARGE have

become ‘mega-consortia’, with almost every cohort study available involved in their analyses.

4.1.2 The CKDGen consortium

The Köttgen *et al* 2013 paper previously referred to several times in this thesis was the work of the Global Urate Genetics Consortium (GUGC), and at >140,000 individuals, still represents the largest GWAS of uric acid to date – though not the most recent, as the Biobank Japan recently published results from their own cohort with > 109,000 individuals and identified several new loci. This will have been in part due to the different ancestry of the cohort (East Asian compared to European) and also due to the use of a denser imputation panel (1000 Genomes phase 1 version 3 compared to the much older and smaller HapMap version 2).

Due to considerable overlap between research interests and membership, the GUGC has been incorporated into the existing CKDGen consortium, which focuses on traits related to kidney function and disease. Serum urate and gout are new additions for their ‘round IV’ analyses – the fourth large set of meta-GWAS run by the consortium. This round aims to be the largest and most efficient yet run, using HRC-imputed data where possible.

An analysis plan was released to collaborators in July 2016, and, shortly after this, I joined the analysis group to co-run the serum urate and gout meta-analyses. The project is nearing completion, with papers for eGFR and UACR (urinary albumin-to-creatinine ratio) both submitted at the time of writing. My contribution to these projects was to run the study-level GWAS for our six cohorts (see Section 4.2.1.2) , and also sharing code I developed for the serum urate meta-analysis (for genetic risk scores and circos plot generation). At time of writing, our paper is under review at *Nature Genetics*.

The analyses performed evolved over time as the results of each step shaped our priorities and new methods were identified or developed. **Figure 29** is an overview of the workflow of the serum urate analyses as it appears in the final paper. My contributions to each section are detailed in Section 4.2 (Methods).

4.1.3 A note on contributions

The collaborative nature of a consortium meta-analysis means that much of the work in this chapter has been done as part of a team, with different people contributing to

different stages. In some places I have detailed work done by others to allow a complete explanation of later work. I have endeavoured to make it clear where my own contributions begin and end.

The bulk of the statistical analysis on this project was performed by myself and Adrienne Tin (Johns Hopkins Bloomberg School of Public Health). At the time of writing, the writing group for our paper comprised myself, Adrienne Tin, Yong Li (University of Freiburg), Karsten Sieber (GlaxoSmithKline, Pennsylvania), Holger Kirsten (University of Leipzig), Matthias Wuttke (University of Freiburg), Mathias Gorski (University of Regensburg), Markus Scholz (University of Leipzig), Adriana Hung (Vanderbilt University), Alexander Teumer (University Medicine Greifswald), Cristian Pattaro (University of Lübeck), Veronique Vitart (University of Edinburgh) and Anna Köttgen (University of Freiburg).

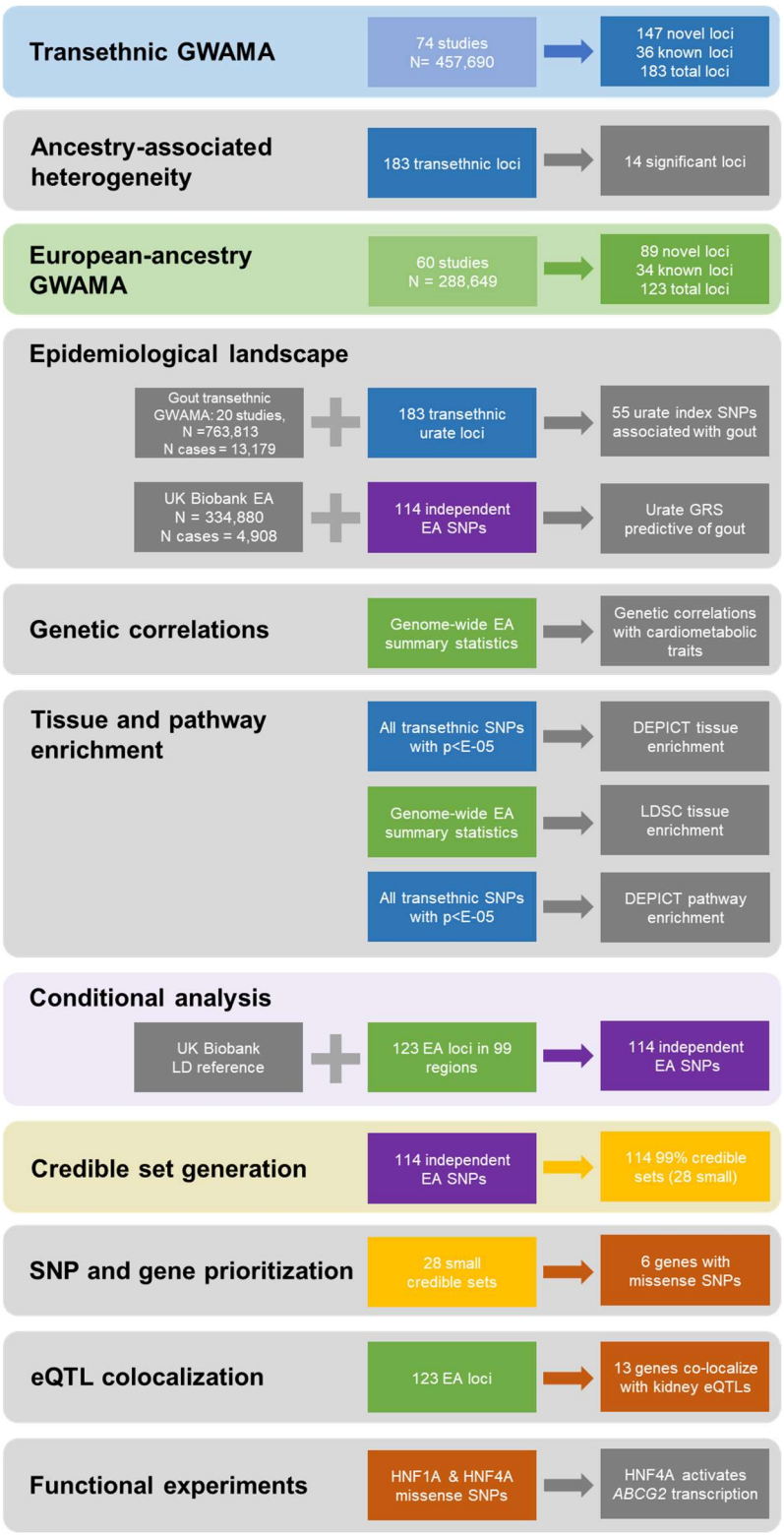


Figure 29 - Overview of the CKDGen serum urate analysis workflow.

Sex-specific and non-European ancestry specific meta-analyses are not shown here, as they did not feed directly into later analyses.

4.2 Methods

4.2.1 Study level analysis & QC

The writing of the analysis plan and phenotype processing pipelines were completed before I joined the CKDGen consortium. I ran the study-level GWAS for our own cohorts, but I was not part of the study level quality control (QC) step. Details are included here for context.

4.2.1.1 Phenotype preparation script

Studies were recruited to the meta-analysis if they had genotype data imputed to the HRC v1.1 (preferred for studies of European ancestry) or 1000G phase 3 version 5 (for studies of non-European ancestry), although those imputed to the 1000G phase 1 version 3 panel were also included in some cases. Phenotypes were processed using an automated script written by Dr Matthias Wuttke and Dr Mathias Gorski that performed all necessary transformations and corrections to produce analysis-ready phenotypes, as well as diagnostic plots and statistics.

4.2.1.2 Study-level GWAS

My contribution to the study-level analysis was to prepare the phenotypes for the CROATIA-Vis, CROATIA-Korcula, CROATIA-Split, ORCADES, VIKING and Generation Scotland: Scottish Family Health Study cohorts and adapt the output of the phenotype generation script for our RegScan GWAS pipeline described in Section 3.2.3, which included the necessary adjustment for relatedness for our cohorts. The summary statistics submitted to the CKDGen meta-analysis were not exactly the same as those analysed in Chapter 3 due to slightly different filtering criteria in the pre-regression stage but were broadly similar. CKD was defined in CKDGen as eGFR < 60 $mL/min/1.73m^2$. Our cohorts did not contribute to the binary phenotypes included due to low case numbers, and our HRC-imputed binary analysis pipeline not being implemented at the time.

The CROATIA-Split study (Split) was not included in the Chapter 3 meta-analysis as it was only imputed to the 1000Gv1p3 panel at the time of analysis. Split is the third cohort in the CROATIA study includes 976 individuals from the city of Split on the Croatian mainland. Unlike Vis and Korcula, it is not an isolate population.

4.2.1.3 Quality Control

Once study-level GWAS results had been returned to the CKDGen analysis group, QC was partially-automated using GWAtoolbox[225] to check for distribution of allele frequencies, inflation of p-values, imputation quality and genotype completeness, all of which can be a sign that a study has not correctly run or filtered the analysis. Metrics were also compared across studies to identify outliers.

4.2.2 Meta-analysis

Meta-analysis was performed by two analysts in parallel to minimise the possibility of human error affecting the results. Dr Adrienne Tin (Johns Hopkins Bloomberg School of Public Health) and I conducted the meta-analyses for serum uric acid, with the additional assistance of Zhi Yu (Johns Hopkins Bloomberg School of Public Health) for the meta-analyses of gout. We used a modified version of METAL (provided by the author) which outputted floating point values to six decimal places, a higher level of precision than is normally available. Meta-analyses were run on the entire combined sample (trans-ethnic), separately by sex (sex-specific) and separately by ancestry (ancestry-specific).

4.2.2.1 Pre-analysis filtering

Study-level results were pre-filtered before meta-analysis to include only biallelic SNPs with a MAC > 10 and an imputation quality score > 0.6 (the precise metric to define this varied depending on the imputation software used by each cohort, but they are sufficiently comparable that a consistent filter can be used).

4.2.2.2 Genomic control

Study-level p-values were corrected for inflation using the λ_{GC} genomic control factor[226]. This optional adjustment is implemented within METAL.

Genomic control is a commonly-used method within the GWAS community as a fast and reasonably effective way to correct inflation of p-values due to population structure. The underlying assumption is that most SNPs are not truly associated with the trait of interest, so the distribution of test statistics should follow a chi-squared distribution with a mean of 1, but population structure can lead to an inflation of the test statistics by a factor λ. This factor can be estimated as median of the χ^2 test statistics divided by 0.456. Dividing the χ^2 values for each SNP by this λ_{GC} factor increases the p-values to compensate for inflation.

Criticisms have been made of the genomic control factor, suggesting both that it is anti-conservative[227] and that it is overly conservative when estimated over large numbers of loci[228]. Studies often apply two levels of GC correction, first at the study level and then again on the meta-analysis results, though this too has been criticised[229], and with a study of this size, inflation is expected from the true positive findings alone under a polygenic trait architecture. However, it remains a commonly-used technique within the community, and its performance and limitations are well understood. Coupled with the study-specific corrections for relatedness and population structure implemented by the study analysts, who are familiar with the requirements of their cohorts (adjustment for PCs, for example or fitting a mixed model with a kinship matrix as I did with our own cohorts), we decided that single genomic control (correction at the study level, but not at the meta-analysis level) was an appropriate method to use in this case.

4.2.2.3 Post-analysis filtering

Post-analysis filters were applied on a per-meta-analysis basis to exclude SNPs present in fewer than half the studies or with a combined MAC < 400. Ancestry-specific meta-analyses were additionally filtered on the heterogeneity I^2 statistic > 95%, to exclude SNPs influenced by a small number of studies within that ancestry group. These filters were selected to reduce the number of single-SNP loci seen in the analysis results – in a sample of this size, a single significant SNP was considered by the analysis group to be likely a false positive.

4.2.2.4 LD-score regression

After analysis, λ_{GC} was calculated on the meta-analysis results to check for inflation. LD-score regression was also used to distinguish between inflation of p-values due to polygenicity and inflation due to confounding[149]. This was calculated using the LDSC software package[148,149]

Briefly, a SNP's LD score is a measure of how much genetic variation is tagged by a SNP. Variants which are in LD with a causal variant will have higher test statistics. SNPs with high LD scores – SNPs that are in LD with a larger number of other SNPs – have a higher probability of being in LD with a causal SNP. Thus, inflation that correlates with LD scores is due to polygenicity – the trait being controlled by large numbers of SNPs with increasingly small effects. Inflation that is uncorrelated with LD score is due to confounding from population structure or relatedness. Regressing

GWAS χ^2 statistics on LD-score (hence the term LD-score regression) allows the contribution of confounding to inflation to be estimated.

4.2.2.5 Locus definition

Loci were defined with the same definition as used in Chapter 3. SNPs with the lowest p-values were defined as index SNPs and any significant SNPs (p-value < 5 x 10^{-8}) within ±500kb assigned to the locus of that index SNP. This process was performed iteratively until all significant SNPs were assigned to loci. Overlapping loci were combined and the index SNP with the lower p-value used as the index SNP for the combined locus

Ancestry-specific loci were defined as loci in which the index SNP did not fall within a locus identified in the transethnic meta-analysis.

4.2.2.6 Sex-specific effect differences

A two-sample t-test was used to compare the effects of SNPs between sexes. T statistics were calculated using the following equation:

$$T = \frac{\beta_F - \beta_M}{\sqrt{{SE_F}^2 + {SE_M}^2}}$$

Where βF and β_M are the regression coefficients for a SNP in females and males respectively, and SE_F and SE_M are the standard errors of the regression coefficient in females and males respectively.

4.2.3 Trans-ethnic meta-regression in MR-MEGA

Heterogeneity in effect sizes between studies can be a sign that no true association between genotype and phenotype exists – this is one reason for running a meta-analysis, as it allows spurious associations in one cohort to be identified as discrepancies. However, heterogeneity can also be real and in a transethnic meta-analysis, can be correlated with ancestry, leading to a loss of power and inability to detect an effect that may be real.

Ancestry-associated heterogeneity can arise because of varying LD patterns between ethnic groups. A SNP in high LD with a causal SNP in one population will have a lower p-value than the same SNP in another population where the LD between the SNPs is lower. In a sufficiently ethnically diverse sample, this effect can be used to fine-map the locus and identify the causal SNP. Heterogeneity can also be due to

environmental factors. A risk factor which modifies the effect of a SNP may have different exposures across populations. Unless this risk factor is explicitly accounted for in the model, it will increase heterogeneity in the effect and reduce the power to detect the association in the meta-analysis. Variation in imputation quality can also drive ancestry-associated heterogeneity. Most meta-analyses use imputation panels to impute all study populations to the same set of SNPs. The performance of the panel depends on how accurately it reflects the study population – a less well-matched panel will lead to lower imputation quality, which will tend to bias allelic effects downwards.

To account for these problems, several software packages have been developed. We made use of MR-MEGA[230] (Meta-Regression of Multi-Ethnic Genetic Association) from the Morris lab, an evolution of the method previously implemented in MANTRA[231]Similar to MANTRA, MR-MEGA uses a matrix of mean pairwise allele frequency differences to quantify genetic similarity between studies. This is constructed by taking the mean difference in allele frequency over all shared SNPs between every pair of studies. While MANTRA uses a computationally demanding Markov chain Monte Carlo implementation of a Bayesian method that fits genetic similarity between studies as a prior, MR-MEGA instead takes principal components of the same matrix of genetic similarity and includes them as covariates in a linear regression framework.

MR-MEGA outputs an estimate of the effect of each SNP after adjusting for the ancestry principal components, as well as a p-value for this effect. Additionally, it provides p-values for ancestry-associated heterogeneity at the SNP ($P_{anc\text{-}het}$) and for residual heterogeneity after ancestry has been corrected for ($P_{res\text{-}het}$). It also provides effect sizes for each of the ancestry principle components, allowing some interpretation of which groups are contributing to the heterogeneity.

Our sample is not sufficiently ethnically diverse to make use of the fine-mapping potential of MR-MEGA, comprising mostly European samples, but accounting for heterogeneity may reveal additional associations that the more transethnic meta-analysis in METAL lacked power to detect. Additionally, the MR-MEGA methodology is more sophisticated than simply including self-reported ancestry group as a covariate: it is able to distinguish between genetically distinct subgroups of cohorts which may share a self-reported ancestry group. For example, a Finnish population would be expected to be quite genetically distinct to an Iberian population.

The MR-MEGA trans-ethnic meta-regression analysis for the CKDGen project is entirely my own work, with the exception of the study-level QC which was shared with the main meta-analysis.

4.2.3.1 Pre-analysis filtering

The same study-specific filters were applied as used in the trans-ethnic meta-analysis (imputation quality > 0.6, MAC > 10). To reduce the influence of very large studies, an additional filter of MAF > 0.0025 was applied. This greatly reduced the amount of noise in the results driven by very rare SNPs that passed the MAC filter through the sheer size of the contributing study.

4.2.3.2 Principal components

Three PCs were fitted, as per the author's recommendation, which proved sufficient to separate the cohorts into self-reported ancestry groups, as well as further separating isolate populations within groups, as shown in **Figure 30**. Due to restrictions of the software, SNPs had to be present in a minimum number of cohorts equal to or greater than the number of PCs plus three. Consequently, any SNPs that were present in five or fewer cohorts were excluded from this analysis.

4.2.3.3 Post-analysis

Results for MR-MEGA were reported both for the index SNPs identified in the transethnic meta-regression; these are included in our paper currently under preparation. MR-MEGA was also run genome-wide. As in the METAL meta-analysis, SNPs were filtered to remove any which were present in fewer than half the cohorts (N cohorts > 37) and assigned to loci as specified in Section 4.2.2.5. If an index SNP was within ±500kb of one of the 183 METAL index SNPs, it was considered to be the same locus as that SNP. If not, the locus was considered unique to the MR-MEGA analysis.

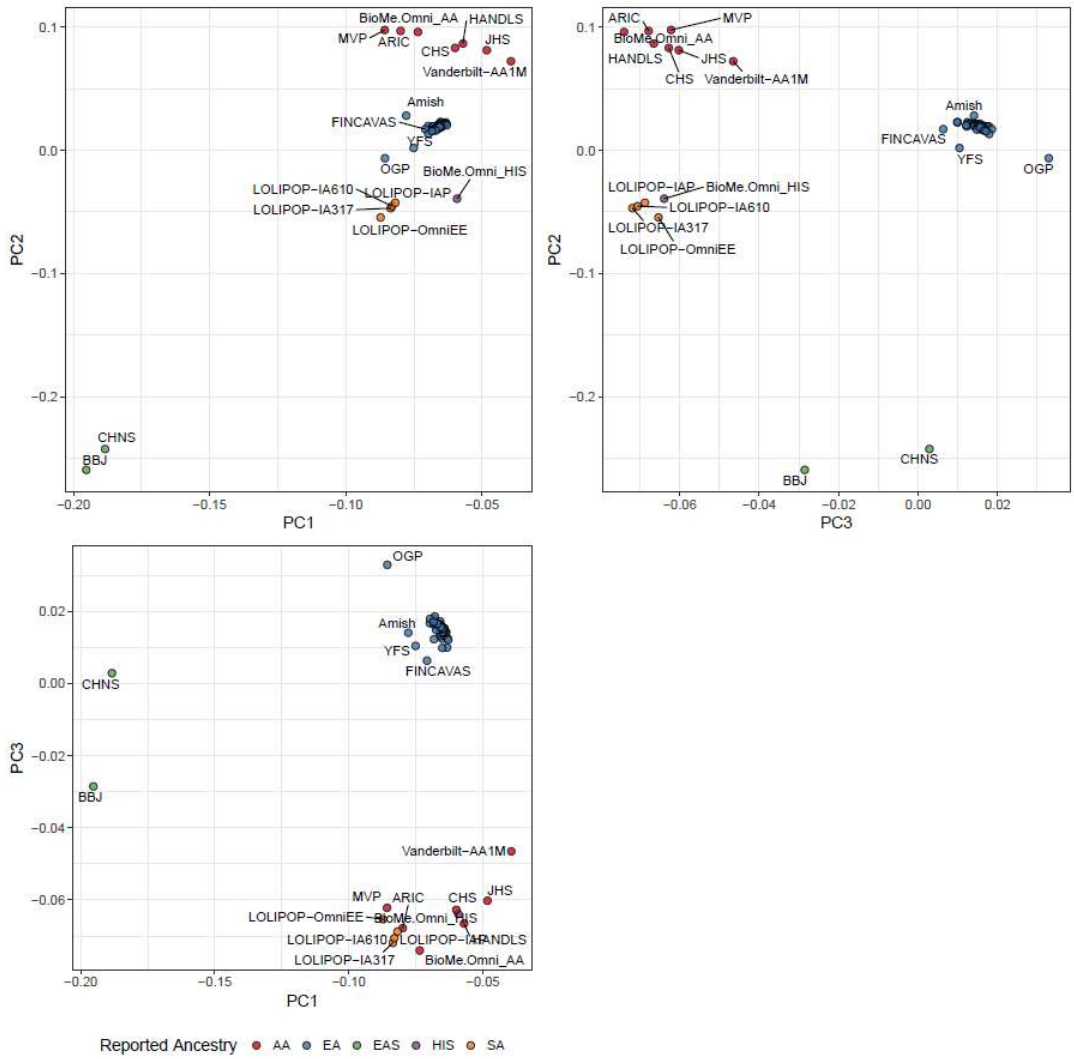


Figure 30 - Principal component plots from MR-MEGA

Each point is one cohort. Point colour corresponds to ancestry reported by cohort, European (EA), African (AA), East Asian (EAS), Hispanic (HIS) or South Asian (SA).

4.2.4 Conditional Analysis in GCTA

Conditional analysis for the CKDGen meta-analysis was performed by Adrienne Tin, but is described here because I made use of the results in downstream analyses.

In GWAS, conditional analysis is a method of determining whether multiple signals are present within a region. In the purest and most straightforward sense, this constitutes including the genotype of the lead SNP in a region as a covariate in the regression model. If a signal is still detected in the region after accounting for the lead SNP, it is demonstrably independent of the conditioned SNP. This is the methodology used in Section 3.2.4.

While this method is theoretically straightforward, it is non-trivial to implement on the scale of a consortium meta-analysis. Each individual cohort in the meta-analysis must re-run the study-level analysis including the genotype of the SNP to be conditioned on, and then all subsequent steps of the meta-analysis must be re-performed, including QC - the burden of time and resources is not dissimilar to that required for the original meta-analysis, which can take months to organise, and this process would have to be repeated until every signal has been accounted for. While this is possible on a small scale where one analyst has access to all the individual level data, and there is only one region of interest, as with the analysis detailed in 3.2.4, it is a prohibitively time consuming and administratively complex process for a GWAS consortium.

To address this problem, a method was developed by the Visscher lab that allowed an approximation to conditional analysis to be performed using only meta-analysis summary statistics – "approximate conditional and joint association analysis"[232], implemented in GCTA as the 'cojo-slct' option[233]. This method requires a reference sample with individual-level genotype data from which estimated LD can be calculated.

The method is a step-wise selection procedure based on p-values – in approximate terms, it uses LD structure in a reference population to generate a joint model of 'conditionally independent' SNPs that together explain the variation in the phenotype, without needing access to the individual level phenotype data. The method iteratively adds the SNP with the lowest conditional p-value to the model until no more SNPs can be added that significantly improve the model fit.

Because the LD structure of the reference population needs to reflect that of the meta-analysis population, we were not able to perform conditional analysis on the trans-ethnic results – no suitable reference population exists. Instead, conditional analysis was performed on the European sub-analysis, with the UK Biobank selected as the reference population from which to estimate LD. After filtering, 13,558 individuals were retained, and 16,969,363 randomly-selected SNPs used to calculate an LD reference panel.

Index SNPs and loci were defined as per Section 4.2.2.5. Neighbouring loci with correlated index SNPs ($r^2 \geq 0.2$) were combined into independent regions. GCTA was then used to identify independent SNPs within each region, using a threshold of $r^2 < 0.01$. For any region with more than one independent SNP, conditional analysis was

conducted using the 'cojo-slct' option in GCTA. These conditionally independent SNPs were used as input in several of the downstream analyses.

4.2.5 Genetic risk score and gout in UK Biobank

To assess the link between our identified serum urate loci and gout, we investigated the relationship between a genetic risk score (GRS) for serum urate and gout status in the UK Biobank cohort[75,234]. This analysis was performed under UK Biobank Projects 19655 and 20272.

The GRS was constructed using the conditionally-independent index SNPs identified in the EA ancestry-specific meta-analysis. The EA index SNPs were used rather than the transethnic firstly to most closely match the genetic background of the cohort, and secondly because the approximate conditional analysis in GCTA was only performed on the EA-ancestry results (see Section 4.2.4).

The cohort was filtered using UK Biobank provided metadata to select only those with genetically defined 'White British' ancestry, remove related individuals with a kinship coefficient greater than 0.0313 and remove any cases of mismatch between self-reported and genomically-inferred sex or sex chromosome aneuploidy.

Gout status was primarily ascertained from self-report at initial clinic visit. Hospital admission records were also used to exclude individuals who developed gout after recruitment, using the ICD10 code that corresponds to gout (M10). This exclusion was done because the 'age' variable used in the model reflected the age of the participant at the time of clinic visit, not at the age at which they had gout. The final analysis dataset comprised 334,880 individuals, of whom 4,908 were gout cases.

The GRS for each individual was calculated as the sum of the additive imputed dosages of the serum urate-increasing index SNP alleles ('risk alleles') weighted by the effect of each allele on serum urate levels as obtained from the EA-specific meta-analysis (of which Biobank UK was not part).

Once GRS had been calculated, the UK Biobank sample was divided into ten bins of equal width between the maximum and minimum GRS values. The lowest bin did not contain any gout cases, which made later analyses nonsensical, so it was merged with the second lowest.

Using the bin with the largest number of individuals as the reference category, a logistic regression model was run regressing gout status on GRS bin, including age

and sex as covariates. This allowed calculation of each GRS bin odds ratios for gout, relative to that of the most common GRS.

Additionally, the GRS was tested for its utility in the prediction of gout status. The dataset was divided into a 90% training sample and a 10% testing sample. Logistic regression models were run in the training set regressing gout status on GRS bin alone (the 'genetic model'), on age and sex alone (the 'demographic model') and on GRS, age and sex together (the 'combined model'). Each model was then used to predict gout status in the testing set, and the performance of the predictor assessed by comparing predicted to true gout status, using the area under the curve (AUC) in a receiver operating characteristic (ROC) curve.

4.2.6 DEPICT pathway analysis

A meta-analysis of this size identifies a large number of significant loci, which are likely to have complex and biologically-meaningful relationships, for example through shared biological pathways or tissue expression. However, manual annotation is time-consuming, so a variety of bioinformatic tools have been developed to aggregate data and automate some of the processing.

We used DEPICT[235] to annotate the results of the meta-analysis. This tool takes GWAS summary statistics as input and outputs a causal gene for each locus, as well as reporting pathways or tissue expression enriched in the results.

I did not run the DEPICT software myself – this was done by Yong Li – but I performed the downstream processing of the pathway enrichment analysis, so the methodology is described here (DEPICT was additionally used to identify tissue expression enrichment, detailed in our paper, but I was not directly involved).

The exact set of SNPs used as input is a decision left to the user, but the developers suggest two runs, one using independent genome-wide significant SNPs ($P < 5 \times 10^{-8}$) and another using all independent SNPs with $P < 10^{-5}$. The final results reported in our paper use all variants from the trans-ethnic meta-analysis with $P < 10^{-5}$ as initial input. We used Plink v1.9[236] to identify independent SNPs in this set using the 'clump' function, which aims to retain only one SNP from each LD block. LD blocks were defined based on $r^2 > 0.1$ in the 1000 genomes phase 1 version 3 data. Summary statistics for these independent SNPs were used as input for DEPICT.

DEPICT uses input SNPs to generate a list of loci, including genes for consideration as potentially causal if they are within $r^2 > 0.5$ of the lead SNP. Loci containing overlapping genes are merged into one locus. The programme then uses a 'guilt-by-association' model to predict the most likely candidate from the genes in the loci, prioritising genes that have similar biological functions across loci.

To do this, DEPICT integrates 14,461 existing gene sets, including manually curated pathways from KEGG[237], Gene Ontology[238] and REACTOME[239], a database of protein-protein interactions[240] and mouse knock-out phenotypes from the Mouse Genome Database[241]. Instead of assigning a binary classification for membership in a gene set, the authors instead created what they refer to as 'reconstituted gene sets' (RGS). Rather than a set including a specified number of genes, each RGS assigns a probabilistic membership value to all genes in the database. These probabilities are assigned based on similarities across gene co-expression data previously published by the developers[242]. The biological function of a gene is thus characterised by its probabilities of membership across all 14,461 RGSs. This step is pre-computed, and the database of RGSs is available for download with the software.

To predict causal genes, DEPICT tests whether any of these RGSs are significantly enriched for genes in these loci. Candidate genes within loci are prioritised based on shared annotations with other loci - for example, if the 'lipid metabolism' pathway is enriched in the results, genes with a high probability of inclusion in that RGS are prioritised within their respective loci.

As well as prioritising candidate genes, DEPICT outputs the results of the RGS enrichment analysis, to provide information on the pathways that appear to be enriched in the GWAS results. I used the output from this in the clustering described in Section 4.2.6.1. DEPICT also uses data from 37,427 microarrays to assess whether genes are highly expressed in any of 209 tissue and cell type annotations, the results of which are described further in our paper.

4.2.6.1 Affinity Propagation Clustering

DEPICT outputs a very large number of enriched gene sets – more than the number of significant loci to interpret in the first instance. This is in part due to overlap between the RGSs – each database may capture the same biological process separately. To reduce the dimensionality of the dataset, I used affinity propagation clustering (AP clustering)[243] to group together RSGs containing overlapping sets of genes.

Affinity propagation clustering (AP clustering) is a clustering algorithm that aims to group together related datapoints and select a single 'exemplar' that best represents the points within the cluster. It does not require pre-specification of the number of clusters, making it ideally suited to this use case, where we have no prior information on the number of pathways to expect to be enriched. It does, however, require an 'input preference' to be specified, which is the propensity of all data points to become an exemplar – this indirectly affects the number of clusters that are identified.

I implemented AP clustering using the R package '*APCluster*'[244]. Reconstituted gene sets identified as significantly enriched in the meta-analysis results (FDR-corrected P-value < 0.01) were clustered based on the similarity of the genes they contained.

DEPICT reports the top ten genes from the meta-analysis assigned to each gene set, along with a z-score representing the probability of that gene's inclusion within the set. This information was converted into a matrix of genes by pathways, where each element contained a z-score. AP clustering was applied to the similarity matrix derived from this data using a tuning parameter of 0.5, as per the package defaults.

To understand the relationships between the exemplars RGSs, the set of exemplars were then used to create a correlation matrix, calculated from Z-score of each gene within the exemplar gene sets. This matrix was converted into a network by taking all exemplar RGSs as vertices and all correlations > 0.2 as edges. The network was plotted in Cytoscape[136].

4.2.7 FUMA gene function annotation

FUMA is a web-based platform that integrates a variety of resources for annotation of GWAS results[245]. Like DEPICT, it aggregates information from a variety of databases, and its primary function is to take raw GWAS output, which can be as little as a list of RSIDs and P-values, and identifying likely causal genes ("SNP2GENE" function). Independent SNPs are assigned to genomic risk loci based on defined LD blocks. SNPs in these blocks are annotated for function consequences, based on gene function, deleteriousness score, potential regulatory function and chromatin state. Functionally-annotated SNPs are then assigned to genes based on their functional consequence on that gene – based on position relative to the gene, eQTL associations and chromatin interaction mapping.

It can then take this gene list and perform enrichment analyses for tissue expression, pathways and gene sets as well as eQTL lookups chromatin mapping ("GENE2FUNC"). It is similar to DEPICT in its aim to aggregate, but it does not introduce the additional complication of the reconstituted gene set method.

Although FUMA is capable of assigning SNPs to genes, the SNP2GENE step is highly customisable, and properly investigating the appropriate parameters to use (for example whether to use only exonic SNPs for positional mapping, or which tissues to restrict eQTL mapping to) would have required more time than was available. Instead, I elected to use only the GENE2FUNC aspect of the FUMA tool on the list of 183 genes identified from our transethnic meta-analysis. This does not provide an entirely fair comparison to DEPICT, as this list of genes is based on the closest gene to each SNP, but it did allow me to gain experience using another annotation tool, and the information on gene-set enrichment allows some comparison to DEPICT. The tool additionally reports tissue enrichment.

4.2.8 Genetic correlations

The genetic correlation analyses described on the CKDGen serum urate paper currently under preparation were run by Yong Li for the sex-combined European-ancestry meta-analysis. All 832 traits in LD-Hub were tested.

In fact, prior to this, I been performing my own genetic correlation analyses on the published results of the Köttgen *et al.* (2013) meta-analysis using cross-trait LD-score regression on LDHub, as described in Section 2.3.7. In addition to performing my own updated genetic correlation analysis on the sex-combined results analysis, I performed additional LD-score regression on the results of the male and female European-ancestry meta-analysis. A two-sample t-test was used to test for significant differences in genetic correlation between sexes, with T statistics calculated with the equation below.

$$T = \frac{r_F - r_M}{\sqrt{{SE_F}^2 + {SE_M}^2}}$$

Genetic correlations were calculated with all traits in the LDHub catalogue except those classified as 'metabolites' or those derived from the UK Biobank. These datasets do contain potentially informative phenotypes, but they are both large and contain many very similar phenotypes (for example, measurements for impedance for

multiple body parts in the UK Biobank) leading to a very large multiple testing penalty on significance for little immediate benefit. In total 128 traits were tested.

4.3 Results

4.3.1 Summary of participants

A summary of the sample sizes for the serum urate and gout meta-analyses is given in **Table 19**, broken down across the five cohort-reported ancestry groups – African ancestry (AA), European ancestry (EA), East Asian ancestry (EAS), Hispanic ancestry (HIS) and South Asian ancestry (SA). Mean serum urate across studies ranged from 4.1 mg/dl to 7.1 mg/dl.

Table 19 – Sample sizes for serum urate and gout meta-analyses

Ethnicity	Serum urate		Gout		
	N	N cohorts	N total	N cases	N cohorts
AA	33,671	7	3,271	282	2
EA	288,649	60	739,330	12,643	17
EAS	125,725	2	-	-	-
HIS	608	1	-	-	-
SA	9,037	4	6,548	253	1
Total	**457,675**	**74**	**749,149**	**13,178**	**20**
Women Total	172,102	66	68,265	438	1
Men Total	174,111	69	84,734	712	1

As is almost always the case in consortium GWAS, the largest and most diverse group was EA, with an order of magnitude more cohorts than any other group. We also have a large sample of EAS, but this is almost entirely composed of the Biobank Japan[76] cohort, which has already published a GWAS of uric acid, limiting our ability to detect new association in this ancestry group[77]. However, taken together, we have over 169,000 non-European ancestry individuals in our analysis, around 37% of the sample. Our transethnic approach should identify signals which are universal across all populations

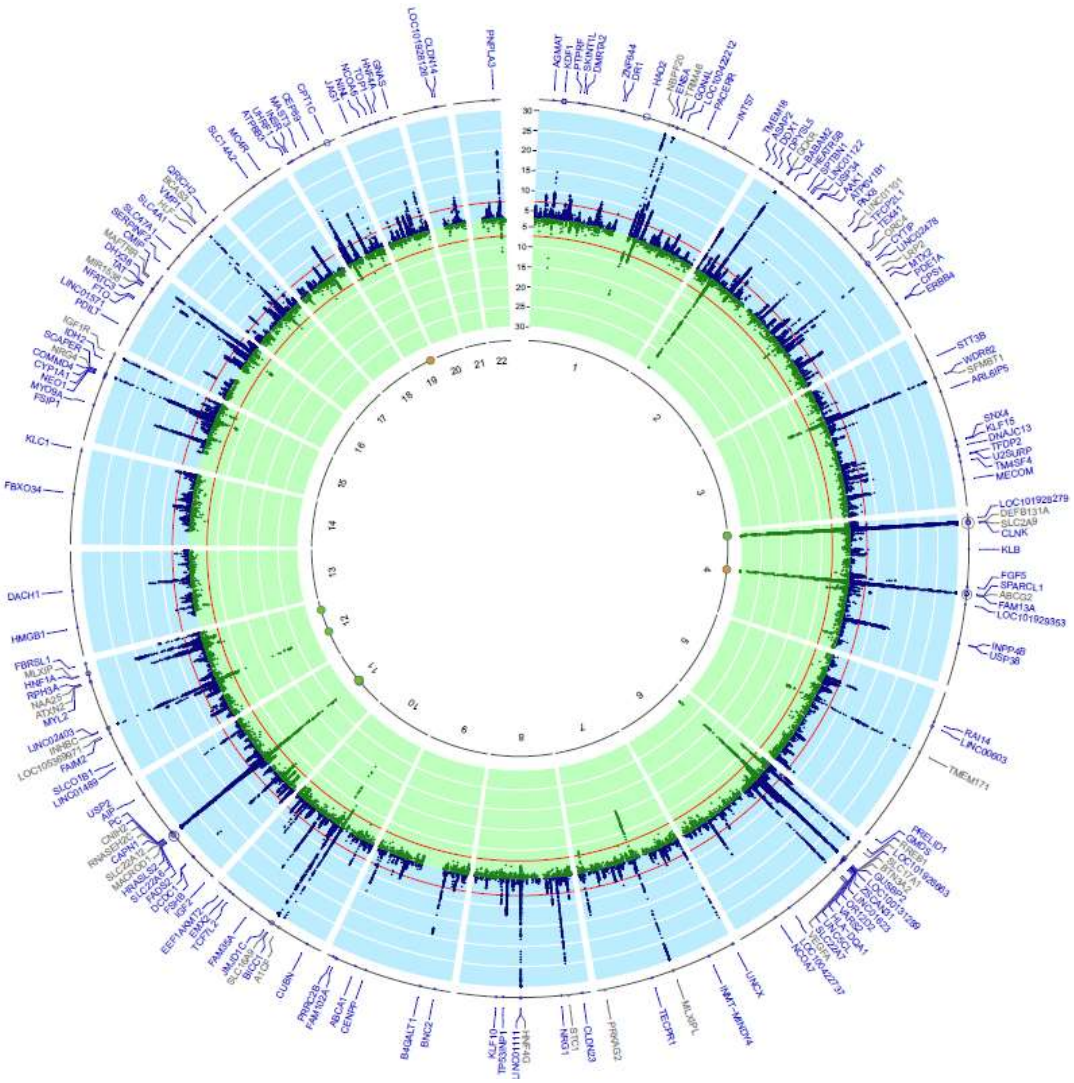
4.3.2 Transethnic meta-analysis

Seventy-four studies were included in the primary uric acid trans-ethnic meta-analysis, with 40,534,360 autosomal SNPs retained after study-level filtering. There was no evidence for confounding due to population stratification – the LD-score regression coefficient was 1.01 and λ_{GC} was 1.04, so no post-meta-analysis correction

of p-values was performed. After post-meta-analysis filtering, 8,249,849 SNPs were retained for downstream analysis.

One hundred and eighty-three loci were identified containing at least one SNP with P-value below the threshold for genome-wide significance ($P < 5 \times 10^{-8}$), spanning a region of maximum 1Mb around the index SNP. Of the 183 loci, 36 contained at least one SNP previously reported as genome-wide significant in a GWAS of serum urate[51,66,74,77,246–249]. The remainder were considered novel. The average absolute effect size was small (mean effect 0.04 mg/dL, SD 0.03), but sizes ranged from 0.017 mg/dL for rs11940694 at the novel *KLB* locus to 0.28 mg/dL at rs3775947 at the well-characterised *SLC2A9* locus. These results are plotted in **Figure 31**, in the blue Manhattan plot track. Index SNPs are fully detailed in **Supplementary Table 4**. Loci were identified in parallel by myself and Adrienne Tin, who also annotated the index SNPs with the name of the closest gene and prepared the original table from which mine is modified.

The heritability and variance explained by these variants were calculated by Adrienne Tin. Together, all 183 index SNPs explained 7.7% of trait variance, compared to just 5.3% when restricting to index SNPs in previously-reported loci. The heritability explained was calculated in the large ARIC study, and the index SNPs were found to explain 17% of the genetic heritability of serum urate. Index variants in the three major loci *SLC2A9*, *ABCG2* and *SLC22A12* alone explained 5%.



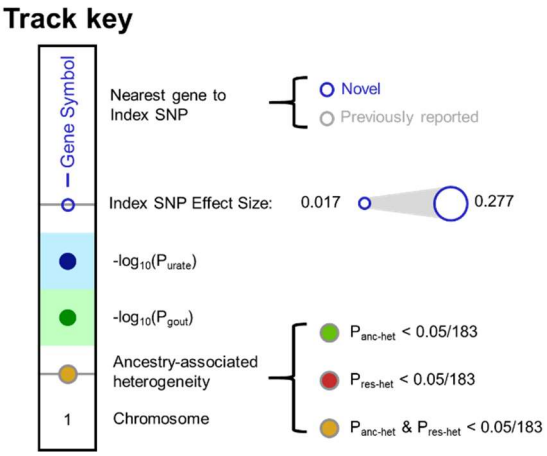


Figure 31 - Circos plot summarising the sex-combined trans-ethnic meta-analysis results.

Serum urate meta-analysis results are shown in the blue Manhattan plot track, gout results in green. The red lines denote the threshold for genome-wide significance.

The track key describes each track from the outermost inwards.

4.3.3 Ancestry-specific analyses

4.3.3.1 European-ancestry

In the EA-specific serum urate meta-analysis, 8,217,339 of the 24,830,632 autosomal SNPs analysed by METAL were retained for downstream analysis, with an LD-score regression intercept of 1.0, indicating no confounding due to population stratification. 123 loci were identified in the EA analysis, of which five loci were not in the trans-ethnic analysis.

4.3.3.2 Non-European ancestry

Ancestry-specific annotation was performed by Adrienne Tin. The results are outlined here to allow comparison to the MR-MEGA results. Five of the fourteen loci identified in the AA analysis were unique to that ancestry group, and seven of the 46 identified in EAS. These unique loci are detailed in **Table 20**.

Table 20 – Ancestry-specific loci not identified in the transethnic meta-analysis.

Ancestry	SNP	Gene Name
EA	rs2480712	*SKI*
EA	rs55781567	*CHRNA5*
EA	rs35396326	*NECTIN2*
EA	rs12037861	*HLX-AS1*
EA	rs98270	*NIPAL1*
AA	rs334	*HBB*
AA	rs7100851	*HABP2*
AA	rs10769187	*PHF21A*
AA	rs2306027	*LRP4*
AA	rs7114004	*LOC101929497*
EAS	rs117247077	*KMT5B*
EAS	rs78863347	*CPT1A*
EAS	rs74896528	*SESN2*
EAS	rs7645142	*EIF5A2*
EAS	rs60808706	*KCNQ1*
EAS	rs35612982	*CDKAL1*
EAS	rs703978	*ZMIZ1*

4.3.4 Trans-ethnic meta-regression in MR-MEGA

4.3.4.1 Trans-ethnic index SNPs

Most transethnic SNPs showed low heterogeneity, indicated by low I^2 values from METAL (median 2%, interquartile range 0-14%). Fourteen of the 183 index SNPs

identified in the primary transethnic meta-analysis showed significant ancestry-associated heterogeneity at a threshold of P < 0.000273 (0.05/183). These were in the *SLC2A9*, *SLC22A12*, *CNIH2*, *CAPN1*, *HRASLS2*, *MACROD1*, *ABCG2*, *CLNK*, *AIP*, *SLC22A6*, *MYL2*, *CPT1C*, *INHBC* and *DEFB131A* loci. These results are displayed on the $P_{anc\text{-}het}$ track in **Figure 31**, and detailed in full in **Table 21**. The most significant ancestry-associated heterogeneity was observed at the *SLC2A9* locus for the index variant rs3775947 ($P_{anc\text{-}het} = 1.5 \times 10^{-127}$). This was reflected in the by effect size differences across the different ancestry-specific analyses (0.34 mg/dL in EA, 0.26 mg/dL in AA, 0.17 mg/dL in EAS, 0.41 mg/dL in HIS and 0.21 mg/dL in SA), and is consistent with known heterogeneity at this locus[250].

4.3.4.2 MR-MEGA-specific loci

MR-MEGA analysis was run across whole genome and, after filtering, loci were identified from the SNP-phenotype association p-values. Nine loci were identified in the MR-MEGA analysis which did not contain an index SNP in the METAL analysis. These are detailed in **Table 22**. Of these, two were overlapping with the locus surrounding a METAL index SNP (i.e. the MR-MEGA and METAL index SNPs were within 1Mb of each other). These were the *ALDH5A1* and *KCNQ1*, both of which were single-SNP loci in MR-MEGA. All loci except *KCNQ1* and *RP11-576N17.4* have significant ancestry heterogeneity, with $P_{anc\text{-}het} < 0.05/9$. The strongest effect was at the *TAPT1-AS1* locus, with an effect size of 1.481 mg/dL. This was also the most significant locus. The locus with the largest number of SNPs was *SLC2A2* with 18, which encodes the GLUT2 glucose transporter, a glucose transporter from the same family as GLUT9. **Table 23** shows the results for these SNPs in the ancestry-specific meta-analysis.

The signals at *KCNQ1* and *ZMIZ* identified in the EAS-specific meta-analysis were also picked up with the MR-MEGA method using the whole sample. None of the five EA-specific or the five AA-specific loci were identified in the MR-MEGA results.

In the case of the AA-specific signal at rs334 at *HBB* (encoding the sickle cell allele) this is a consequence of post-analysis filtering on the number of cohorts. The index SNP is detected as significant by both METAL and MR-MEGA in the trans-ethnic analysis, but is present in only eight cohorts, leading to its exclusion from the results. This is not the case for the other ancestry-unique SNPs. **Supplementary Table 5** contains the unique MR-MEGA index SNPs with the cohort constraint relaxed.

Table 21 – METAL index SNPs with significant ancestry-associated heterogeneity ($P_{anc\text{-}het}$ < 0.05 / 183)

rsID	Novel?	Chr.	Position	Gene	Novel	Function	EA/NEA	EAF	Effect (mg/dL)	SE	p-value	I2 (%)	$P_{anc\text{-}het}$	$P_{res\text{-}het}$
rs6820627		4	9,491,205	*DEFB131A*		upstream	A/G	0.071	-0.100	0.011	4.66E-21	28.6	2.54E-02	3.20E-05
rs3775947		4	9,995,240	*SLC2A9*		intron	T/C	0.691	0.277	0.003	0.00E+00	90	1.10E-109	1.50E-127
rs12504795	*	4	10,499,344	*CLNK*	Y	intron	T/C	0.741	0.075	0.004	4.25E-101	51.4	2.74E-07	1.70E-18
rs74904971		4	89,050,026	*ABCG2*		intron	A/C	0.196	0.217	0.004	0.00E+00	73.7	1.58E-25	3.90E-25
rs148838714	*	11	62,732,352	*SLC22A6*	Y	intron	A/G	0.066	-0.074	0.009	2.14E-16	46.6	8.48E-05	1.40E-09
rs143825439	*	11	63,319,993	*HRASLS2*	Y	near-gene-3	T/G	0.054	-0.224	0.012	1.23E-76	76.4	1.69E-17	4.80E-27
rs1006207		11	63,849,812	*MACROD1*		intron	T/C	0.563	-0.047	0.004	3.54E-41	64.7	4.72E-14	3.20E-26
rs531763		11	64,352,063	*SLC22A12*		intergenic	A/G	0.565	-0.116	0.004	1.58E-246	83.7	2.30E-55	1.50E-67
rs34888828	*	11	64,968,104	*CAPN1*	Y	intron	A/G	0.110	-0.057	0.006	1.45E-23	75	6.91E-27	1.10E-49
rs4073582		11	66,050,712	*CNIH2*		intron	A/G	0.310	-0.041	0.004	5.41E-28	79.2	4.14E-37	1.00E-66
rs11227805	*	11	67,246,757	*AIP*		upstream	T/C	0.180	-0.027	0.005	8.72E-09	49.4	4.61E-06	4.00E-15
rs73119306		12	57,826,982	*INHBC*		near-gene-5	A/G	0.786	0.071	0.004	7.45E-65	33.5	3.69E-03	6.80E-06
rs17550549	*	12	111,357,471	*MYL2*	Y	intron	T/C	0.144	-0.035	0.005	5.18E-11	33.9	3.29E-03	4.60E-09
rs62128132	*	19	50,217,955	*CPT1C*	Y	downstream	T/C	0.967	-0.115	0.014	1.99E-15	72.7	7.42E-16	3.00E-07

Table 22 – MR-MEGA loci which did not contain a METAL index SNP

rsID	Chr.	Position	Gene	Function	EA/NEA	EAF	N	N cohorts	N eth.	N SNPs	Effect	SE	P	$P_{anc\text{-}het}$	$P_{res\text{-}het}$	METAL Effect	METAL I2 (%)
rs3774046	3	170,737,003	*SLC2A2*	Intron	A/G	0.840	457,616	74	5	18	0.050	0.048	1.84E-09	4.40E-04	0.161	0.0207	25
rs697238	10	80,947,668	*ZMIZ1*	Intron	T/G	0.386	456,290	73	5	5	-0.026	0.034	2.46E-09	1.37E-03	0.0291	-0.016	32.3
rs17325213	4	11,955,802	*TAPT1-AS1*	Intergenic	T/C	0.056	239,512	48	4	4	-1.481	0.258	3.59E-09	5.51E-08	0.378	-0.0301	42.4
rs11601310	11	48,085,189	*PTPRJ*	Intron	A/G	0.236	443,108	69	5	1	-0.011	0.039	8.60E-09	3.40E-05	0.412	0.0179	22.8
rs73728140	6	24,507,003	*ALDH5A1*	Intron	A/G	0.954	445,653	69	5	1	0.064	0.056	1.03E-08	2.88E-09	0.514	-0.0073	32.3
rs144074240	12	67,657,757	*GGTA2P*	Downstream	T/C	0.061	330,887	70	4	3	0.121	0.127	1.60E-08	1.33E-03	0.968	0.0385	0
rs60808706	11	2,857,233	*KCNQ1*	Intron	A/G	0.160	446,270	72	5	1	-0.013	0.055	1.67E-08	0.0586	0.753	0.0253	0
rs1346144	4	79,625,361	*RP11-576N17.4*	Intergenic	A/G	0.631	422,278	72	5	1	-0.186	0.052	2.24E-08	7.53E-03	0.893	-0.0172	0
rs508926	1	28,578,825	*RP5-1092A3.5*	Downstream	G/A	0.280	447,631	73	5	1	0.034	0.034	4.69E-08	2.03E-05	0.191	0.0131	28.2

Table 23 – Lookups of MR-MEGA unique index SNPs in ethnic-specific meta-analyses.
SNPs significant at a nominal threshold of P < 0.05 are highlighted in yellow. Genome-wide significant SNPs are highlighted in green.

SNP	*Gene*	Eth.	EA/NEA	EAF	Effect	Std. Err.	P-value	N	I2
rs3774046	*SLC2A2*	AA	A/G	0.895	0.061	0.022	5.60E-03	33,671	0
		EA	A/G	0.838	0.009	0.005	9.34E-02	288,575	12.8
		EAS	A/G	0.822	0.044	0.008	1.42E-08	125,725	0
		SA	A/G	0.890	0.011	0.032	7.32E-01	9,037	0
rs697238	*ZMIZ1*	AA	T/G	0.426	-0.033	0.013	1.24E-02	33,670	20.3
		EA	T/G	0.412	-0.008	0.004	3.01E-02	288,614	30.8
		EAS	T/G	0.322	-0.032	0.006	2.09E-07	125,725	27.1
		SA	T/G	0.368	-0.027	0.021	1.97E-01	9,037	0
rs17325213	*TAPT1-AS1*	AA	T/C	0.017	-0.073	0.072	3.11E-01	28,882	48.4
		EA	T/C	0.063	-0.023	0.010	1.74E-02	256,970	41.3
		EAS	T/C	0.040	-0.046	0.067	4.91E-01	9,037	0
rs11601310	*PTPRJ*	AA	A/G	0.816	0.074	0.017	6.55E-06	33,671	0
		EA	A/G	0.199	0.024	0.005	7.70E-07	274,067	5.8
		EAS	A/G	0.173	-0.007	0.008	3.69E-01	125,725	0
		SA	A/G	0.125	-0.006	0.030	8.38E-01	9,037	0
rs73728140	*ALDH5A1*	AA	A/G	0.719	-0.036	0.015	1.36E-02	33,670	0
		EA	A/G	0.972	-0.032	0.012	7.26E-03	276,614	0
		EAS	A/G	0.973	0.088	0.018	7.26E-07	125,725	0
		SA	A/G	0.969	0.055	0.059	3.52E-01	9,037	0
rs144074240	*GGTA2P*	AA	T/C	0.011	-0.138	0.066	3.62E-02	32,653	0
		EA	T/C	0.069	0.038	0.008	9.95E-07	288,590	0
		SA	T/C	0.055	0.147	0.045	1.13E-03	9,037	0
rs60808706	*KCNQ1*	AA	A/G	0.197	0.012	0.017	4.98E-01	33,671	0
		EA	A/G	0.057	0.010	0.009	2.48E-01	278,593	0
		EAS	A/G	0.390	0.035	0.006	5.63E-09	125,725	67.9
		SA	A/G	0.028	-0.091	0.071	2.01E-01	9,037	0
rs1346144	*RP11-576N17.4*	AA	A/G	0.238	0.026	0.019	1.63E-01	33,670	0
		EA	A/G	0.646	-0.016	0.004	1.22E-04	278,592	0
		EAS	A/G	0.609	-0.022	0.006	1.60E-04	125,725	0
		SA	A/G	0.538	-0.014	0.020	4.94E-01	9,037	0
rs508926	*RP5-1092A3.5*	AA	A/G	0.507	-0.013	0.013	3.38E-01	33,670	0
		EA	A/G	0.670	0.010	0.004	1.17E-02	278,591	15.5
		EAS	A/G	0.888	0.049	0.010	2.13E-07	125,725	0
		SA	A/G	0.760	-0.024	0.024	3.08E-01	9,037	0

4.3.5 Sex-stratified analysis

Four loci were detected in females that were not present in males or the sex-combined analysis – i.e., there was no significant SNP in either GWAS within a 1Mb window centred on the index SNP. These loci are detailed in **Table 24**. All are small loci (< 10 SNPs) with P-values relatively close to the threshold for significance. They are not explicitly reported in the CKDGen paper, due to limited space. Of these SNPs, only rs75523587 was reported in PhenoScanner as being strongly associated with another trait, namely heel bone mineral density (BMD) in the UK Biobank ($P = 6.47 \times 10^{-16}$).

The p-values for the formal test of effect size difference between sexes are displayed as in a Manhattan plot in **Figure 32**. Significant differences ($P_{diff} < 5 \times 10^{-8}$) were identified for SNPs in *SLC2A9* and *ABCG2*, both of which have been previously reported. Additional SNPs with suggestive differences are listed in **Table 25**.

Table 24 – Female-specific loci for serum urate.

Index SNP	Chr.	Position	A1/A2	Locus	Freq1	Effect	Std. Err.	P-value	Locus size	I^2 (%)	Known associations
rs494889	1	234851020	C/T	*RP4-781K5.7*	0.431	-0.028	0.0049	1.586E-08	1	0.0	-
rs2588983	10	63566778	A/G	*RP11-491H19.1*	0.1812	0.032	0.0058	3.749E-08	6	12.1	
rs75523587	10	134413583	T/A	*INPP5A*	0.2107	0.033	0.0058	1.206E-08	2	0.0	Heel BMD, hypertension (UKBB)
rs12790943	11	120058623	C/T	*TRIM29*	0.4279	-0.033	0.0054	1.34E-09	6	22.4	-

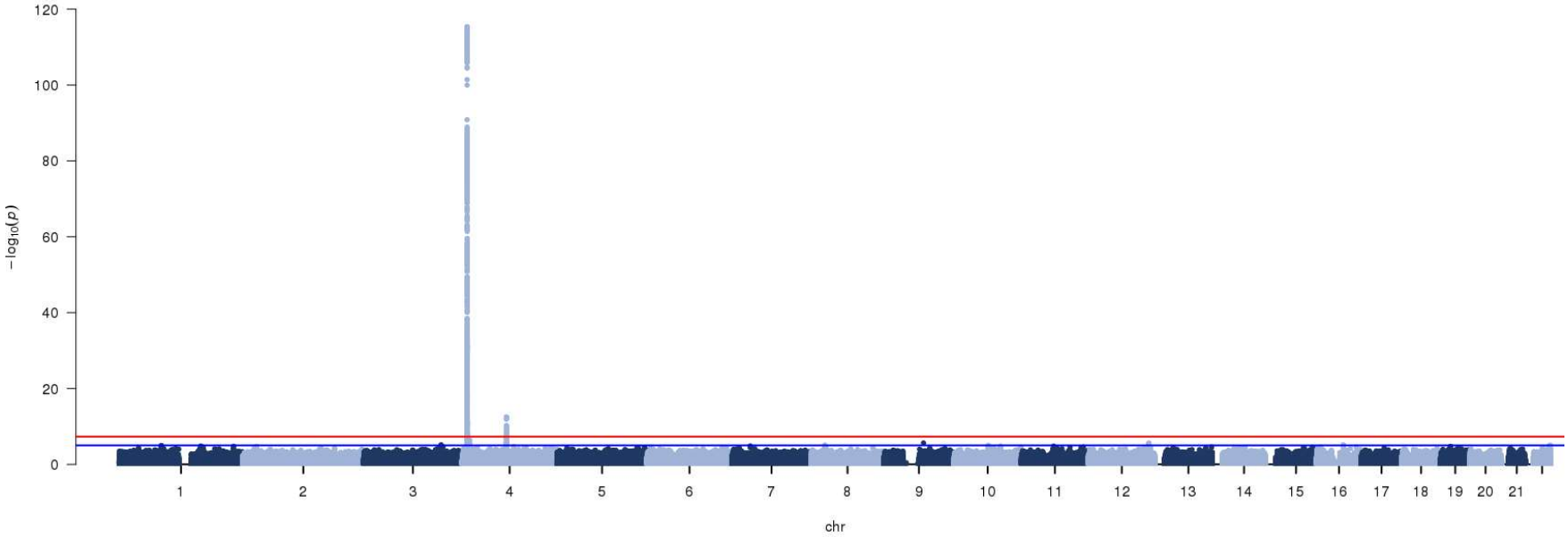


Figure 32 - Manhattan plot of the p-values for effect size differences between females and males.

Blue horizontal line denotes suggestive significance ($P = 10^{-5}$), red denotes genome-wide significance ($P = 5 \times 10^{-8}$).

Table 25 – SNPs with suggestive sex-effect differences ($P_{diff} < 10^{-5}$)

Index SNP	Chr.	Position	A1/A2	Locus	Effect (male)	Effect (female)	SE (male)	SE (female)	Sex difference P-value
rs358716	3	155,526,790	T/C	*AC104472.3*	-0.019	0.015	0.006	0.005	7.32E-06
rs17246501	4	9,985,710	A/C	*SLC2A9*	0.083	0.234	0.005	0.004	0.00E+00
rs3815493	4	15,512,390	A/G	*CC2D2A*	0.023	-0.017	0.006	0.005	4.78E-07
rs2199936	4	89,045,331	A/G	*ABCG2*	-0.231	-0.166	0.007	0.006	3.06E-13
rs7002838	8	27,758,391	C/G	*SCARA5*	0.036	-0.036	0.012	0.011	8.07E-06
rs60866856	9	79,751,986	A/C	*VPS13A*	-0.076	0.090	0.027	0.023	2.32E-06
rs1904619	10	69,048,533	T/C	*CTNNA3*	-0.037	0.059	0.016	0.014	9.54E-06
rs2244608	12	121,416,988	A/G	*HNF1A*	-0.002	-0.034	0.005	0.005	2.42E-06
rs11647020	16	53,823,990	T/C	*FTO*	-0.007	0.030	0.006	0.006	7.36E-06
rs184568395	22	50,720,874	A/C	*PLXNB2*	0.115	-0.070	0.032	0.026	9.13E-06

Table 26 – Lookup of Chapter 3 novel SNPs in CKDGen meta-analyses

Analysis	rsID	Chr.	Position	Locus	EA/NEA	N	EAF	Effect	SE	P-value	N. cohorts	I^2 (%)
All	rs577353818	4	190,587,150	*RP11-462G22.1*	T/C	55,994	0.001	0.198	0.231	0.3899	1	0
All	rs573624409	11	72,799,547	*FCHSD2*	A/G	37,219	0.004	-0.452	0.106	1.91E-05	7	66.6
Female	rs112029032	8	43,054,647	*HGSNAT*	G/A	44,699	0.010	-0.021	0.059	0.7272	7	54.3
Female	rs1876025	11	4,615,987	*OR52I1*	A/C	5,546	0.050	0.131	0.069	0.05836	7	0
Female	rs576082236	14	103,468,041	*CDC42BPB*	G/A	31,407	0.003	0.056	0.134	0.6762	1	0
Male	rs139498948	12	79,289,619	*SYT1*	G/A	40,498	0.049	-0.011	0.049	0.8142	15	57.8
GS Tay.	rs75869162	5	16,617,922	*FAM134B*	G/A	204,884	0.003	0.061	0.058	0.2912	17	63.8
GS Tay.	rs141208451	11	45,538,920	*RP11-430H10.4*	G/A	101,215	0.004	0.111	0.071	0.1208	10	76.6
GS Tay.	rs187171029	19	53,599,256	*ZNF160*	A/T	221,502	0.005	0.086	0.037	0.02066	24	59

4.3.6 Lookup of Chapter 3 SNPs

The index SNPs for the novel loci identified in Sections 3.3.2.2 and 3.3.3 were looked up in the results of the corresponding CKDGen meta-analyses. The results are shown in **Table 26**.

Of these index SNPs, only rs573624409 at *FCHSD2* in the sex-combined analysis was significant after multiple testing correction (two tests for All, three for female, one for male and three for GS Tayside), though heterogeneity at this locus was relatively high (I^2 = 66.6%) despite the SNP only being present in the meta-analysis in seven cohorts (of which only one was a cohort from the Chapter 3 meta-analysis, ORCADES, the others having not passed the pre-meta-analysis filtering).

The forest plot for this SNP is shown in **Figure 33**, with details for each cohort in **Table 27**. These show that direction is consistent across most cohorts, with the strongest signal coming from ORCADES. Despite this being the smallest cohort contributing to this SNP, the minor allele frequency is highest here, and the imputation quality is higher than any other cohort, which suggest this signal may be real. This may be an example of a rare allele being enriched in a population isolate, making it possible to detect in a much smaller sample.

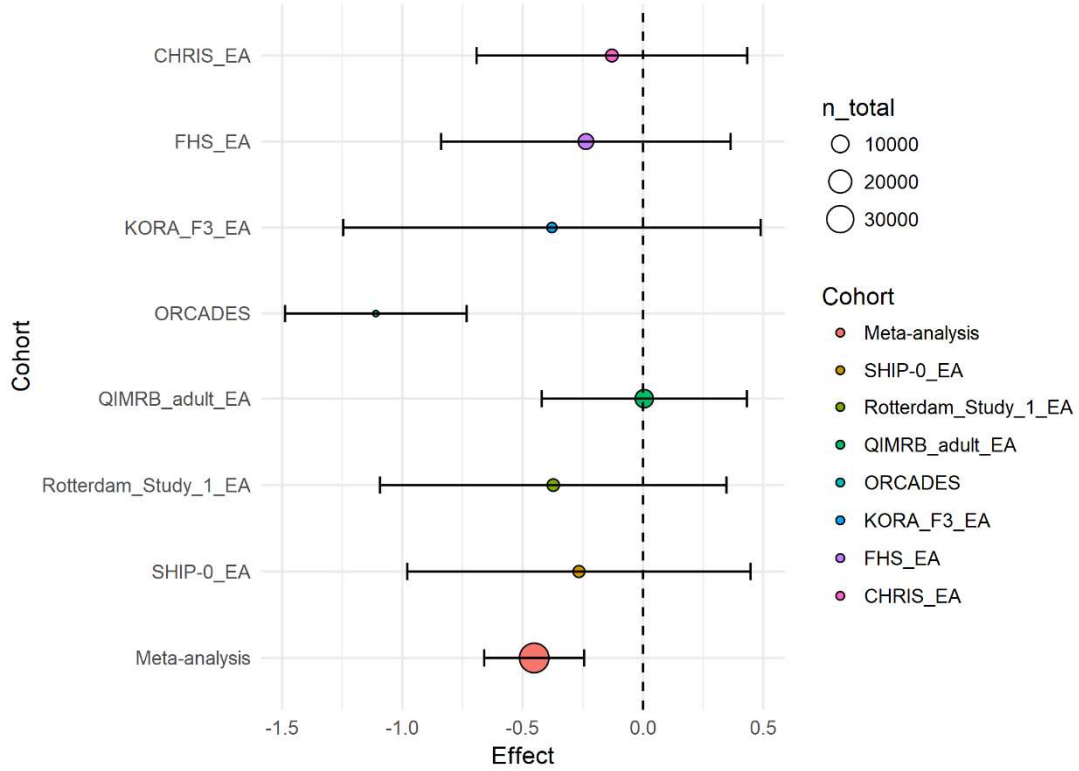


Figure 33 - Forest plot for rs573624409 in CKDGen cohorts.

Table 27 – Study-specific GWAS summary statistics for rs573624409.

Cohort	MAF	MAC	n	Imputation quality	Effect	Std. Err.	P-value
CHRIS (EA)	0.0022	20.14	4,661	0.721	-0.129	0.287	0.653
FHS (EA)	0.0017	25.87	7,699	0.662	-0.237	0.307	0.440
KORA F3 (EA)	0.0018	10.73	2,996	0.675	-0.378	0.443	0.393
ORCADES	0.0084	33.75	2,003	0.894	-1.110	0.193	8.05E-09
QIMRB adult (EA)	0.0021	46.69	11,389	0.669	0.006	0.218	0.980
Rotterdam Study 1 (EA)	0.0018	15.48	4,415	0.793	-0.373	0.367	0.310
SHIP-0 (EA)	0.0019	15.49	4,056	0.696	-0.266	0.364	0.465
Meta-analysis	0.0040	168.15	37,219	-	-0.452	0.106	1.91E-05

4.3.7 Genetic risk score and gout in UK Biobank

4.3.7.1 Gout Odds Ratios

Gout prevalence increased across the serum urate GRS bins, ranging from 0.1% in the lowest category (3.61-4.17 mg/dl) to 12.9% in the highest category (6.15 - 6.44 mg/dl, **Table 28**, left panel, **Figure 34a**). The most common risk score category was 4.74 - 5.02 mg/dl, with almost a third of the sample falling within this range. This was used as the reference category for expressing age- and sex-adjusted odds ratios (OR)

of gout (**Table 28**, right panel). ORs ranged from 0.09 (95% CI 0.02 - 0.37, P = 7.8 x 10^{-4}) in the lowest category to 13.6 (95% CI 7.2 - 25.7, P = 1.4 x 10^{-15}) in the highest category, corresponding to a >100-fold range (**Figure 34b**).

A sizeable proportion of the UK Biobank, 3.46%, has a greater than three-fold increased risk for gout (OR > 3) compared to the most common risk category.

4.3.7.2 Gout prediction

Gout risk prediction models were built by regressing gout status in the UK Biobank dataset on the GRS alone ("genetic model"), on age and sex ("demographic model"), and finally on the GRS as well as on age and sex ("combined model") in a training sample consisting of 90% of the individuals. These models were then used to predict gout status in the remaining testing set and compared to the actual gout status. The area under the curve (AUC) of the corresponding Receiver Operating Characteristic (ROC) curves showed the genetic model to be a moderately accurate predictor of gout status (AUC = 0.68), weaker than the demographic model (AUC = 0.79). The combined model led to a statistically significant increase in predictive ability (AUC = 0.84, DeLong's test Z = -8.43, p-value < 2.2 x 10^{-16}). These results are displayed in **Figure 34c**.

Table 28 – Summary of serum urate GRS bin demographics and logistic regression model.

				Gout ~ age + sex + GRS		
GRS bin range	**Individuals**	**Gout cases**	**Gout prevalence**	**OR**	**95% CI**	**P-value**
3.61 < GRS <= 4.17	2,095	2	0.10%	0.093	0.023 - 0.371	7.80E-04
4.17 < GRS <= 4.46	15,743	61	0.39%	0.363	0.28 - 0.47	1.67E-14
4.46 < GRS <= 4.74	57,758	321	0.56%	0.528	0.466 - 0.599	1.87E-23
4.74 < GRS <= 5.02	105,697	1,097	1.04%	Reference		
5.02 < GRS <= 5.3	97,471	1,622	1.66%	1.616	1.495 - 1.746	9.61E-34
5.3 < GRS <= 5.59	44,513	1,280	2.88%	2.890	2.662 - 3.138	6.75E-141
5.59 < GRS <= 5.87	10,246	423	4.13%	4.240	3.775 - 4.763	4.68E-131
5.87 < GRS <= 6.15	1,264	90	7.12%	7.713	6.121 - 9.717	2.83E-67
6.15 < GRS <= 6.44	93	12	12.90%	13.586	7.161 - 25.777	1.41E-15

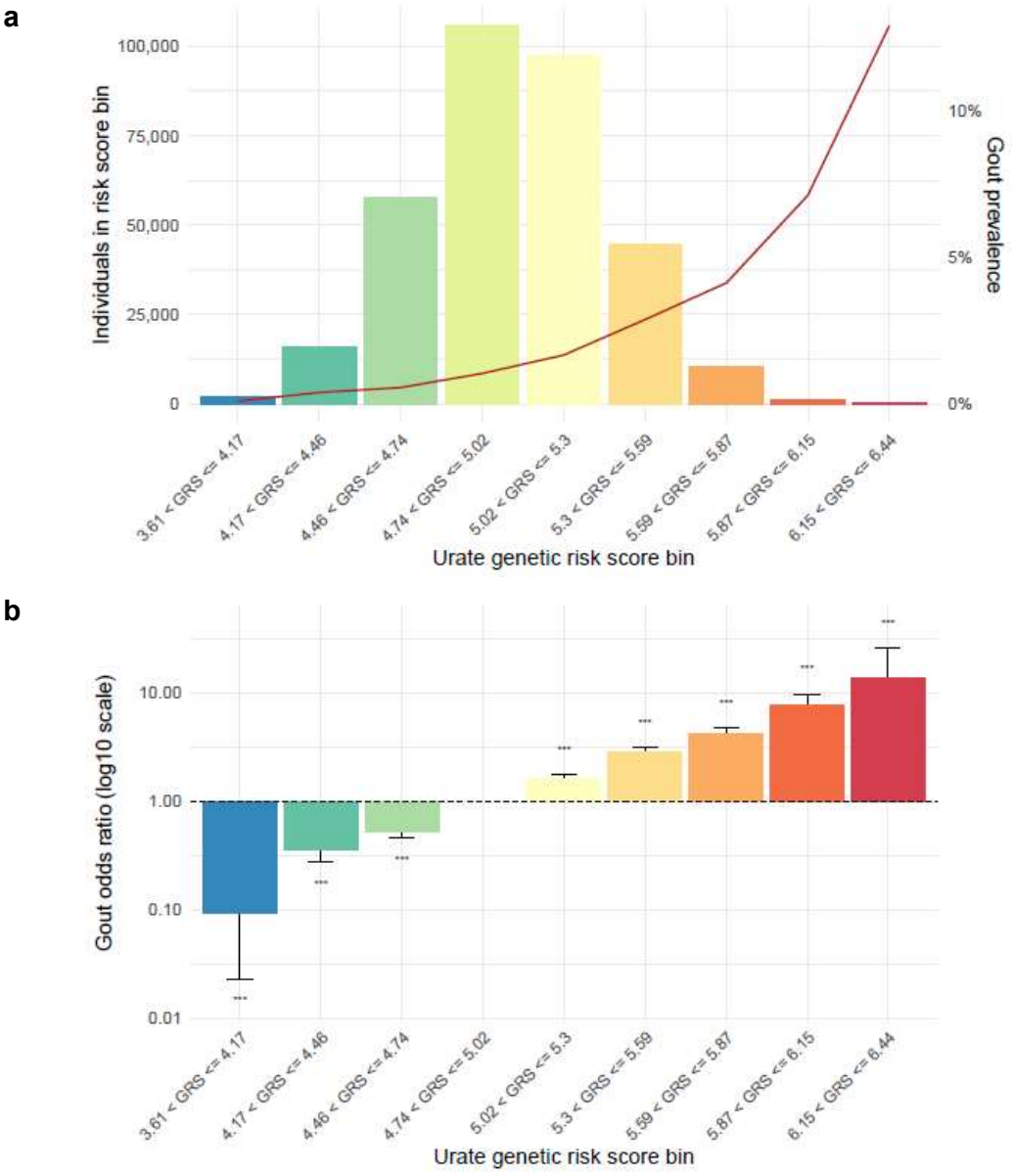


Figure 34 - Serum urate genetic risk score and gout prediction

a Histogram of serum urate GRS in the UK Biobank. Gout prevalence within each GRS bin is plotted on the secondary axis; **b** Gout odds ratios for GRS bins, adjusted for age and sex; **c** Comparison of the ROC curves of demographic (age + sex), genetic (GRS only) and combined (GRS + age + sex) gout predictor models.

4.3.8 DEPICT Pathway Analysis

DEPICT pathway analysis identified 383 significantly enriched reconstituted gene sets (FDR P-value < 0.01), containing 867 genes. The most commonly seen gene was *NPHS2*, which appeared in 46 sets, while 302 genes appeared in only one set – interestingly *SLC2A9* was among this number, appearing only in the "FOXO1 PPI Subnetwork" set. *ABCG2* appeared in three sets: "Renal Tubular Necrosis", "TGF-Beta Receptor Binding" and "Abnormal Embryonic Growth/Weight/Body Size". *SLC22A12* appeared in 41.

It is a noteworthy demonstration of the limitations of pathway analysis methods that although urate-related gene sets exist within the DEPICT database – GO terms alone include several – the only reconstituted gene set with a urate-related name was "GO:0046415: urate metabolic process", which had a nominal P-value of 0.06 and contained *NAT8* (7.8), *SLC7A13* (7.3), *SLC22A11* (7.0), *SLC22A24* (6.3), *ENSG00000204872* (6.0), *PKHD1* (5.8), *CUBN* (5.7), *ACMSD* (5.5), *ENSG00000223985* (5.4) and *SLC22A6* (5.3) as its top ten genes.

Affinity propagation clustering of these 383 reconstituted gene sets identified 57 exemplar gene sets. Constructing a correlation network between these exemplars revealed a large group of inter-correlated gene sets related to kidney development, morphology and function (**Figure 35**, **Supplementary Table 6**).

Table 29 – DEPICT Exemplar Gene Sets.

Gene set source refers to the data source of the original gene set used to construct the DEPICT RGS. The P-value is the nominal P-value before multiple testing correction for enrichment of the gene set in the GWAS results. Cluster size refers to the number of RGS summarised by this exemplar.

Exemplar Gene Set	Gene Set Source	P-value	Cluster Size
ABCC9 PPI	Protein-Protein Interaction	0.000369	7
Abnormal Bone Ossification	Mouse Phenotype	0.000442	5
Abnormal Glucose Homeostasis	Mouse Phenotype	0.000151	10
Abnormal Kidney Physiology	Mouse Phenotype	1.48E-06	8
Abnormal Long Bone Epiphyseal Plate Morphology	Mouse Phenotype	0.000451	3
ACVR1B PPI	Protein-Protein Interaction	0.000266	11
Apical Part of Cell	Gene Ontology	3.17E-08	6
Cell-Substrate Junction	Gene Ontology	7.93E-05	6
CRTC1 PPI	Protein-Protein Interaction	3.81E-05	23
Decreased Circulating Cholesterol Level	Mouse Phenotype	0.000164	1
Decreased Embryo Size	Mouse Phenotype	2.47E-09	11
Decreased Urine Osmolality	Mouse Phenotype	1.68E-10	4
Electron Carrier Activity	Gene Ontology	6.32E-05	6
Embryonic Morphogenesis	Gene Ontology	8.93E-05	6
Enlarged Kidney	Mouse Phenotype	9.01E-07	6
ENSG00000215320 PPI	Protein-Protein Interaction	4.32E-05	4
FHL2 PPI	Protein-Protein Interaction	1.86E-06	3
FLT1 PPI	Protein-Protein Interaction	0.000119	3
Glucose Catabolic Process	Gene Ontology	0.000124	4
HEBP1 PPI	Protein-Protein Interaction	8.48E-07	9
Hexose Metabolic Process	Gene Ontology	5.34E-07	3
Hydronephrosis	Mouse Phenotype	9.75E-07	3
Increased Circulating Aldosterone Level	Mouse Phenotype	5.63E-05	2
Increased Circulating Potassium Level	Mouse Phenotype	3.52E-05	1
Increased Glomerular Capsule Space	Mouse Phenotype	1.38E-05	4
Increased Urine Calcium Level	Mouse Phenotype	1.79E-06	4
Increased Urine Sodium Level	Mouse Phenotype	2.44E-07	4
ITGA5 PPI	Protein-Protein Interaction	0.000361	4
Kidney Cortex Atrophy	Mouse Phenotype	6.84E-06	3
Kidney Development	Gene Ontology	2.61E-06	4
Liver Development	Gene Ontology	0.000237	6
Mesangial Cell Hyperplasia	Mouse Phenotype	5.31E-07	4
Microsome	Gene Ontology	5.01E-05	2
NCOA3 PPI	Protein-Protein Interaction	2.28E-06	24
Negative Regulation of Growth	Gene Ontology	4.75E-05	28

Exemplar Gene Set	Gene Set Source	P-value	Cluster Size
NR3C1 PPI	Protein-Protein Interaction	0.000102	10
Nuclear Hormone Receptor Binding	Gene Ontology	2.11E-05	5
Partial Prenatal Lethality	Mouse Phenotype	4.14E-07	16
PDGFC PPI	Protein-Protein Interaction	4.31E-06	12
PLG PPI	Protein-Protein Interaction	4.93E-05	10
Ppara Activates Gene Expression	REACTOME Pathway	6.83E-05	3
Regulation of Carbohydrate Metabolic Process	Gene Ontology	3.07E-05	3
Regulation of Organ Morphogenesis	Gene Ontology	9.18E-06	6
Respiratory Tube Development	Gene Ontology	0.000304	2
Response to Hypoxia	Gene Ontology	6.67E-05	2
Response to Nutrient Levels	Gene Ontology	6.53E-05	12
Response to Peptide Hormone Stimulus	Gene Ontology	6.26E-05	10
SERPINE1 PPI	Protein-Protein Interaction	8.79E-05	3
Slc-Mediated Transmembrane Transport	REACTOME Pathway	0.000137	4
Small Molecule Catabolic Process	Gene Ontology	0.00034	2
Steroid Hormone Receptor Activity	Gene Ontology	7.49E-05	3
Thoracic Vertebral Transformation	Mouse Phenotype	0.000106	4
Transforming Growth Factor Beta Receptor Cytoplasmic Mediator Activity	Gene Ontology	2.71E-05	12
Tube Development	Gene Ontology	3.91E-08	15
UBE2I PPI	Protein-Protein Interaction	2.21E-05	8
Wnt-Activated Receptor Activity	Gene Ontology	4.37E-05	3
YWHAE PPI	Protein-Protein Interaction	0.000168	6

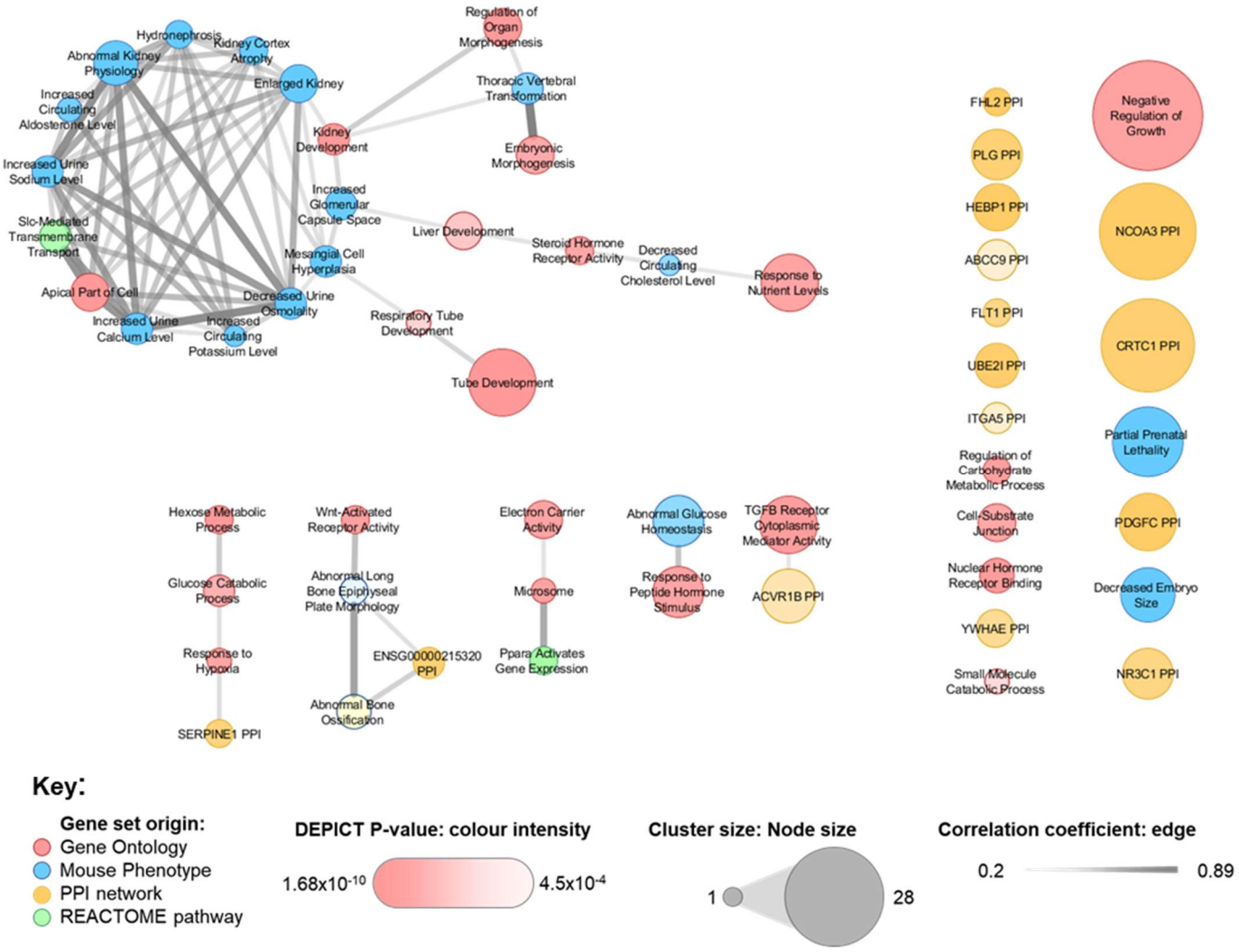


Figure 35 - Correlation network of exemplar gene sets from DEPICT pathway analysis.

Nodes represent exemplar gene sets, edges represent a Spearman's correlation of r > 0.2 between the Z-scores of the top 10 genes in each set.

4.3.9 FUMA gene function annotation

Of the 183 gene names uploaded to FUMA GENE2FUNC, 15 were not recognised in the database, largely uncharacterised non-coding RNAs. FUMA reports enrichment for a range of gene sets, too many to easily summarise, but a few are noteworthy. The strongest enrichment for GO biological processes – one of the datasets included in DEPICT and used to generate reconstituted gene sets – is "Urate Metabolic Process" (adjusted p-value = 2.64×10^{-10}, *SLC16A9*, *SLC22A12*, *GCKR*, *SLC2A9*, *ABCG2*, *SLC17A1*). This gene set was not even nominally significant in DEPICT (see Section 4.3.8) and the top ten genes assigned to it did not overlap with the set reported for this term in FUMA. The next most significant is "Lipid Metabolic Process", containing 23 genes ($P_{adj} = 1.34 \times 10^{-6}$).

A heatmap (generated by FUMA directly) of gene expression across 53 tissue types from GTEx v6 is shown in **Figure 36**, and shows a group of genes highly expressed only in liver, as well as a smaller group expressed highly across a range of brain tissues. Differentially-expressed gene (DEG) analysis (**Figure 37**) in which pre-calculated sets of DEGs were tested for enrichment in the set of genes identified in the GWAS showed significant enrichment for genes differentially expressed in Kidney Cortex, Liver and Subcutaneous Adipose Tissue, with significant up-regulation DEG sets in Kidney Cortex and Liver.

These results support tissue enrichment analyses performed by Adrienne Tin and Yong Li in both DEPICT and stratified LD-score regression (not detailed here) for our CKDGen paper, which identified enrichment for kidney and liver in the results of the EA meta-analysis.

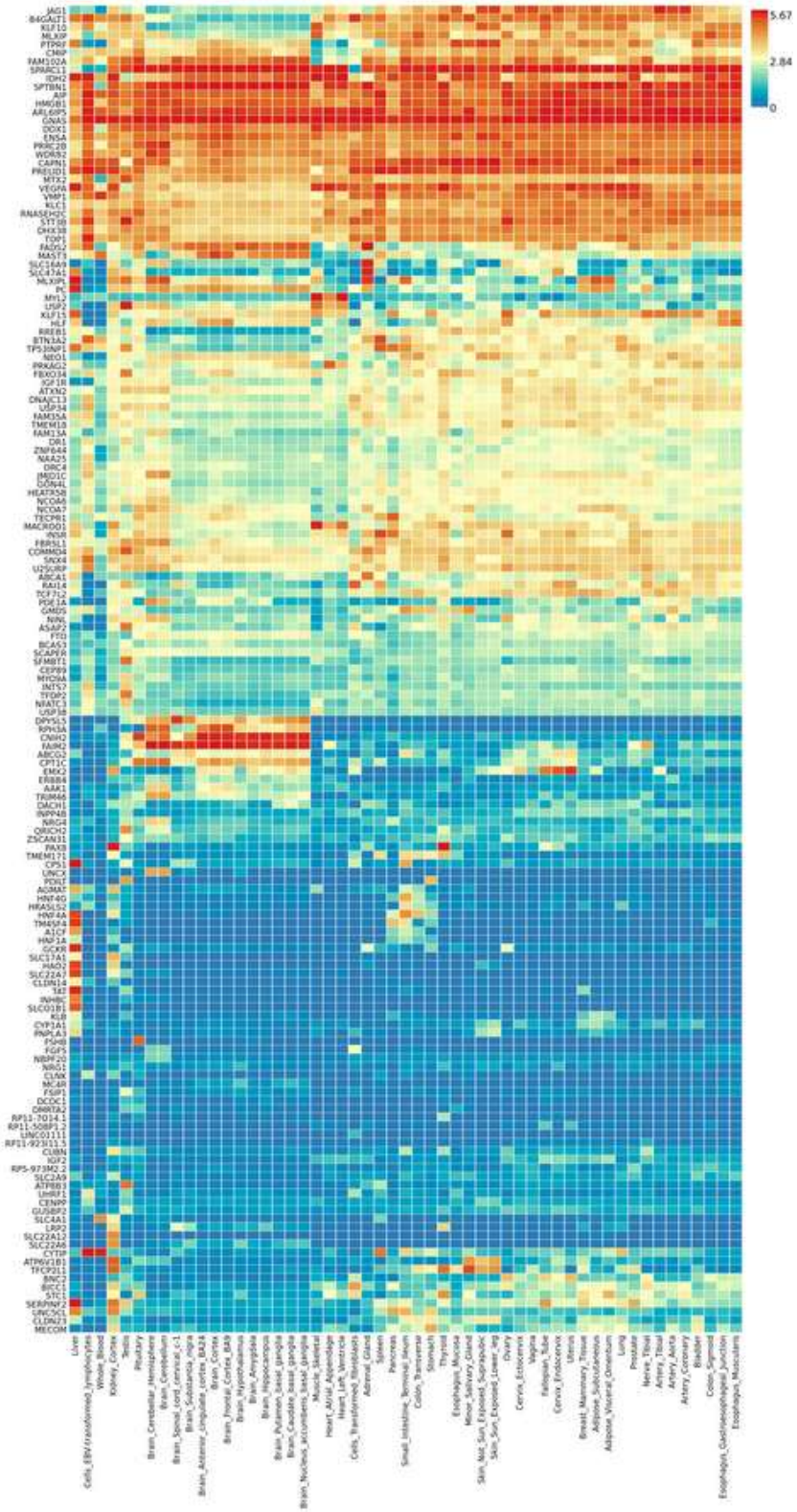


Figure 36 – FUMA-generated heatmap of GTEx gene expression for the serum urate loci.

Cell colour is average log_2 transformed expression per tissue type. Cells are clustered based on FUMA's inbuilt clustering

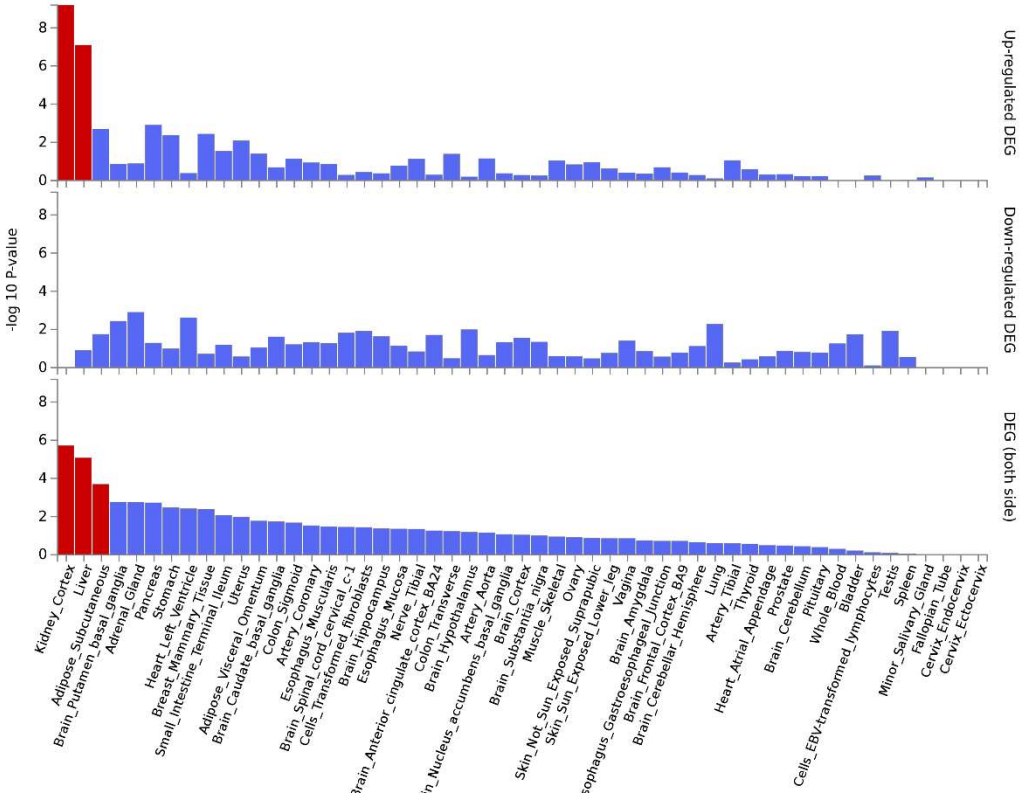


Figure 37 – FUMA-generated plot of differentially-expressed genes by tissue type. Based on enrichment of input genes in pre-calculated sets of DEGs per tissue type. Red bars denote significant enrichment (Bonferroni-corrected P-value).

4.3.10 Genetic correlations

Of the traits tested, 214 had significant genetic correlations with serum urate ($P < 0.05/832$). The strongest positive genetic correlations were with serum urate itself (in previously published GWAS), with values around 1, and with gout. There were also positive strong correlations with metabolic syndrome components including triglycerides in small HDL ($r_g = 0.50$), HOMA-IR (a measure of insulin resistance, $r_g = 0.49$) and fasting insulin levels ($r_g = 0.45$), as well as CKD and cardiometabolic risk factors including obesity, waist circumference, body fat, and type 2 diabetes. Strong negative correlations were observed with a variety of HDL cholesterol-related measurements. **Figure 38** shows all genetic correlations with an absolute r_g value of > 0.35.

Figure 38 – Genetic correlations between serum urate EA and LD-Hub traits

All traits with absolute r_g > 0.35 are shown. Brackets contain the PMID of the GWAS summary statistics used in cross-trait LD-score regression, or else UKB to indicate the UK Biobank.

4.3.10.1 Sex-stratified genetic correlations

4.3.10.1.1 Genetic correlation between sexes

The genetic correlation between serum urate in females and serum urate in males was also calculated using LDSC. The correlation was high (r_g = 0.875, SE_{rg} = 0.047, P = 2.7715 x 10^{-77}) but as seen in **Figure 38**, for the overlap should be 1 or higher for

a trait correlated against itself. This is consistent with the existence of sex-specific genetic variants controlling serum urate levels.

To test for additional undiscovered sex-specific loci, I excluded the two loci with known sex effects, *SLC2A9* and *ABCG2*, by removing all SNPs within 1Mb of the gene boundaries (defined as taken as 4:9,772,777-10,056,560 and 4:89,011,416-89,152,474 respectively). This resulted in a stronger genetic correlation, as expected (r_g = 0.9094, SE_{rg} = 0.0301). However, this value is still smaller than 1, which may reflect additional unidentified heterogeneity between males and females that is below the genome-wide threshold for significance in GWAS.

4.3.10.1.2 Genetic correlations with other phenotypes

Fifty-three of the 128 traits tested showed significant genetic correlation with serum urate in males or females (**Supplementary Table 7**). Of these, none showed significant differences between sexes at an FDR-corrected threshold of Q < 0.05. Nine traits were nominally significant before multiple testing correction (P < 0.05, **Table 30**, **Figure 39**).

Table 30 – Traits with a nominally significant difference in genetic correlation between sexes.

* denotes significant genetic correlations at Q < 0.05. Q-values for sex difference are all < 0.05 and are not shown.

	Trait	Category	r_g Male	r_g Female	Q-value Male	Q-value Female	Difference Score	$P_{sex\text{-}diff}$
More positive in males	Age at Menopause	Reproductive	0.015	-0.115	8.04E-01	1.68E-02*	-2.331	0.020
	Former vs Current smoker	Smoking	-0.050	-0.260	6.06E-01	1.95E-03*	-2.225	0.026
	Years of schooling	Education	-0.062	-0.146	3.26E-02*	2.78E-05*	-2.133	0.033
More positive in females	Childhood obesity	Anthropometric	0.102	0.235	6.70E-02	2.71E-05*	1.973	0.049
	HbA1C	Glycaemic	-0.005	0.168	9.54E-01	2.36E-02*	2.065	0.039
	Coronary artery disease	Cardiometabolic	0.106	0.220	4.01E-03*	1.01E-05*	2.068	0.039
	BMI	Anthropometric	0.196	0.330	1.47E-05*	8.94E-10*	2.102	0.036
	Type 2 Diabetes	Glycaemic	0.141	0.329	2.57E-02*	4.16E-05*	2.102	0.036
	Waist to hip ratio	Anthropometric	0.186	0.329	2.71E-05*	6.78E-09*	2.193	0.028

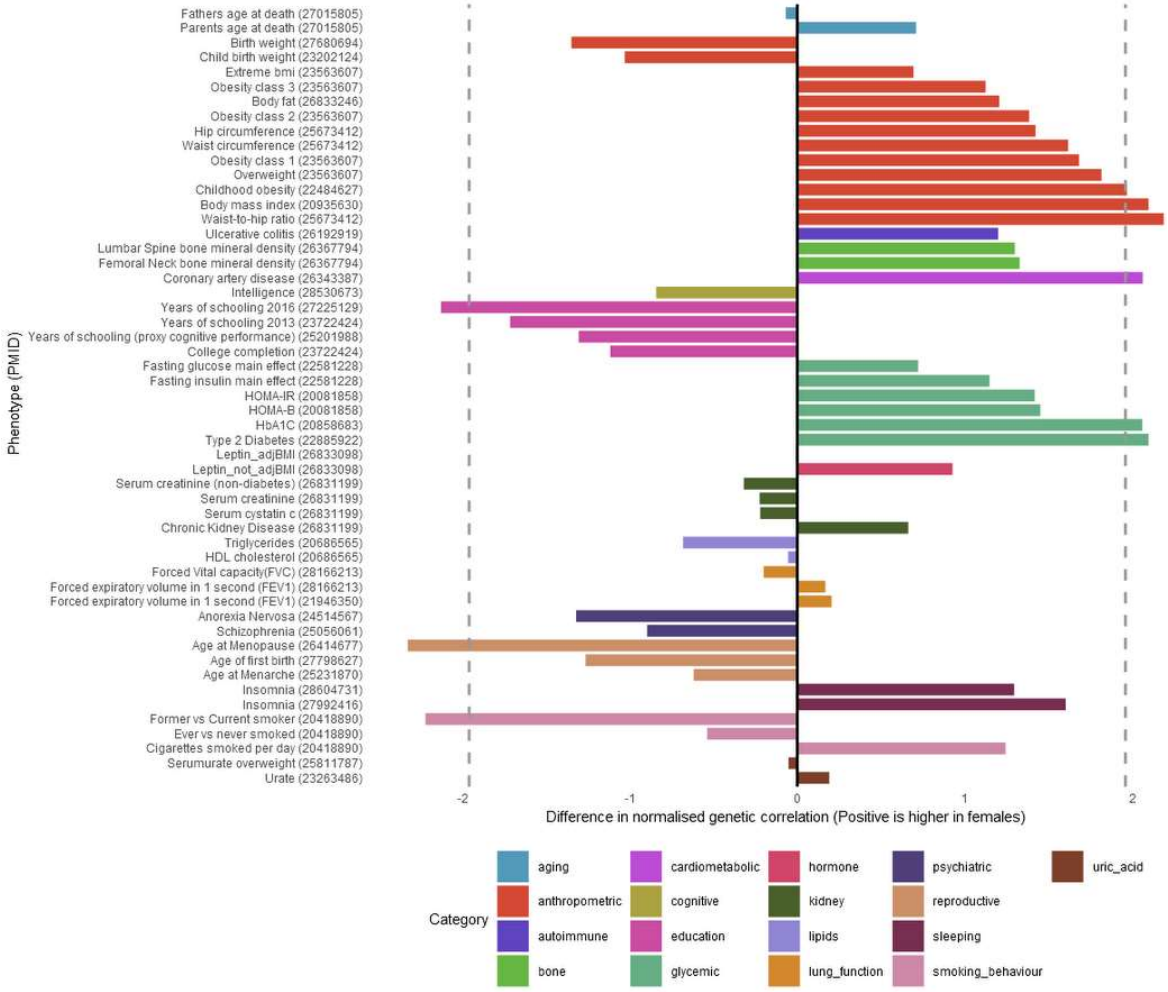


Figure 39 - Comparison of genetic correlations with serum urate between males and females.

Positive values indicate stronger genetic correlations in females, negative indicate stronger in males. Dashed lines indicate the threshold of nominal significance ($P < 0.05$).

4.4 Conclusions

4.4.1 Meta-analysis

This meta-analysis is the largest yet run on serum urate, and represents a four-fold increase on the discovery sample from Köttgen *et al*.[74]. Of the 183 loci reported, 147 were considered novel – not previously reported in GWAS of serum urate[74,77]. This is in part due to the increased density of the new HRC imputation panel, but also reflects the dramatic increase in power that comes from the increased sample size.

However, it must be noted that the definition of novel locus is quite simplistic, being based purely on distance to previous loci, without taking LD structure into account – by necessity, since the transethnic nature of the dataset means LD structure is highly variable between cohorts. Some of the novel loci are clustered around known strong serum urate hits – in particular, the cluster around *SLC22A12* on chromosome 11. If a full conditional analysis (performed at the study level across all ethnic groups and meta-analysed) were performed in this region, or the predefined region of 1Mb per locus was relaxed, it is likely that some of the associations would disappear. However, the majority of novel loci are spaced across the genome, with at least one hit on every autosome, so even a conditional analysis or wider locus window would still leave a very large number of novel associations.

Step-wise conditional analysis on individual study genotype data would have a prohibitively high administrative burden in a meta-analysis such as this, but the UK Biobank – a single cohort with a sample size that exceeds even that of this combined meta-analysis – would be an ideal dataset for testing the independence and fine structure of these loci. It is, of course, a European ancestry cohort, and so might not reflect the variation in non-European populations, but as a single cohort, one analyst could perform a full step-wise conditional analysis – including SNPs one at a time as covariates in the regression model, and joint analyses of local variant effects (including testing for epistatic interactions) – without assumption of a pattern of LD as required in a consortium meta-analysis. Unfortunately, the release of the biochemistry data for the cohort has been delayed, and there is no linkage yet to NHS blood test results, meaning no measures of serum urate are available.

Statistical fine mapping and gene-expression co-localisation analyses performed by the analysis group (not detailed here) were used to prioritise likely causal variants in

a number of loci, including prioritising the *SLC17A4* gene in the 6p22.2 region, which contains four SLC17 family transporter genes.

However, this fine mapping is not without potential flaws. In particular, the credible set generation, where statistical analysis is used to predict a set of SNPs that likely contains the causal SNP, is strongly dependent on the p-values as input. This means that the inclusion of a SNP is sensitive to the contribution and sample size of the combined populations that include it. Specifically, in the *PDZK1* locus, a strong case for a causal variant had been made from experimental work at rs1967017, which modified an HNF4A binding site upstream of the gene[251]. However, this variant is not included in the credible set in our analysis, while a variant in strong LD with it, following 1000G-UK reference samples, is. This could be explained by the fact that due to uneven genotyping arrays and imputation quality across cohorts meta-analysed the selected adjacent variant had a higher sample size, and thus a lower p-value.

This discrepancy was identified by the group in the very late stages of preparation of the paper, and impact of equal contribution of cohorts to region analysed on the credible set results will be examined before final publication. Though unfortunate, this does highlight a limitation of consortium analysis – that decisions must be made as a group, and changes in direction are much more difficult to implement. The conclusions of an analysis are always partially dependent on the methods used to achieve them.

The trend within the GWAS field is towards publishing full results in public repositories such as dbGaP[143], and our meta-analysis will be no exception. This will allow our results to be used in future analyses by researchers across the world. Because of the increasing availability of GWAS summary statistics, the development of methods that can utilise these results without needing individual genotype and phenotype data is an area of active development. LDSC and MR-base are two noteworthy tools using summary statistics which have appeared just within my brief tenure as a PhD candidate, and doubtless more will be developed over time. By making our raw results available to other researchers, we allow them to be re-analysed and integrated into new tools, ensuring our findings will not remain static, instead leading to increasingly accurate insights into the biology of serum urate regulation in the future.

Given the flaws highlighted above, an even better policy might be to publish individual cohort-level GWAS results, allowing future researchers the same level of flexibility and detail as we have had, and allowing others to test our choices of methods and parameters. Unfortunately, this is an unlikely prospect. The additional burden of

hosting the data notwithstanding, individual cohorts would have to consent, and many may have to withhold this to ensure compliance with their ethical agreements with their participants.

4.4.2 Trans-ethnic meta-regression in MR-MEGA

4.4.2.1 Trans-ethnic index SNPs

Fourteen of the transethnic index SNPs were identified as having significant ancestry-associated heterogeneity. Rates of renal urate clearance are known to vary in different populations, specifically being reduced in Polynesian women[252] and Maori men[253], which has been suggested to be due to founder effects increasing the frequency of deleterious alleles in *SLC2A9*[250]. We detected both ancestry-associated and residual heterogeneity at this locus. This suggests that while mean allele frequency differences between populations explains some of the heterogeneity, either there are more granular differences in allele frequency between populations at this locus that are not captured by our three PCs, or that there is another factor driving the heterogeneity of effect – perhaps an environmental interaction that is not related to ancestry.

Heterogeneity has been reported for gout at *ABCG2*, where the apparent causal SNP in Europeans has not been found to have an effect in Maori, but this may have been driven by the frequency of the causal allele being relatively low in this group[254]. *SLC17A1* has been reported as consistently associated with gout despite varying allele frequencies across populations[250], and this is reflected in our results – there is no significant heterogeneity detected at this locus in MR-MEGA or METAL.

Of the remaining loci, *DEFB131A* and *CLNK* are clustered within a 1Mb region around *SLC2A9* on chromosome 4. Seven more are within a 4.5Mb window containing *SLC22A12*, a region where heterogeneity between populations has not been previously reported. The remaining three are dispersed: *INHBC* and *MYL2* on chromosome 12 and *CPT1C* on chromosome 19.

Of the eight loci with the smallest association p-values using MR-MEGA, six had significant heterogeneity (*SLC2A9*, *ABCG2*, *SLC22A12*, *CLNK*, *GCKR* and *HRASLS2*). Similarly, five of the six loci with the largest absolute effects have significant heterogeneity (*SLC2A9*, *HRASLS2*, *ABCG2*, *HAO2*, *SLC22A12* and *CPT1C*). This co-occurrence of strong GWAS effects and detectable ancestry-associated heterogeneity suggests that our findings are at least partially affected by sample size, presumably due to the relatively small number of non-European cohorts.

While the Biobank Japan sample is large, it is one single cohort, and the MR-MEGA method is only able to account for heterogeneity between cohorts, not within them – this has to be accounted for at a study level. Additional and larger non-European cohorts would also allow us to utilise MR-MEGA's fine-mapping capabilities. This takes advantage of LD structure variation between populations to narrow down the region containing the causal SNP.

The bias towards Europeans in GWAS cohorts has been a characteristic of the field since the very beginning. This is because European ancestry cohorts were the first to be collected and genotyped – a consequence of the countries which have funded population genetics research having interest in their own populations first. Generation Scotland, for example, is funded in part by the Scottish Government as a research resource. While a cohort of 20,000 individuals from an African country might have been be a highly useful resource from a scientific perspective, it lacks the political relevance of a study of the local population, making it unlikely to secure government support. This state of affairs is rapidly changing as it becomes more apparent that diverse samples are needed, but as it can take years to gather the data for a cohort, progress is slow.

4.4.2.2 Loci unique to MR-MEGA analysis

Despite the predominantly European sample, MR-MEGA identified nine loci with small but significant association signals which did not overlap with the loci from the METAL analysis. These are reported in the CKDGen paper despite the comparatively low p-values for the index SNPs as several of them remain of interest.

4.4.2.2.1 *SLC2A2*

The most significant of these loci is *SLC2A2*, with an index SNP P-value of 1.84 x 10^{-9}. The locus contains 18 significant SNPs and the effect of the index SNP, while modest, is estimated as over twice as large by MR-MEGA compared to METAL (0.050 compared to 0.0207). Several SLC family proteins have been associated with serum urate levels in previous GWAS, making this a promising candidate. Variation in this gene has been associated with response to metformin in obese individuals, a medication used to treat type 2 diabetes. Like other SLC2A family genes, *SLC2A2* encodes a glucose transporter, GLUT2 – a high-capacity, low-affinity glucose transporter which plays a role in glucose uptake in the liver and in glucose reabsorption in the kidney proximal convoluted tubule[255]. It is also expressed in the intestine and the central nervous system. Knockouts of GLUT2 in mouse has shown

it to be essential for glucose reuptake in the kidney, but dispensable for uptake in the gut[255]. Xenopus oocytes injected with *SLC2A2* cRNA were not found to have significant serum urate uptake, in contrast to those injected with *SLC2A9* cRNA[57]. If GLUT2 has a true effect on serum urate levels, it is likely to be mediated through its effect on glucose regulation.

Lookup of the *SLC2A2* index SNP in the ancestry-specific meta-analyses shows that it has its strongest effect in AA but is only nominally significant due to the low sample size. It is actually genome-wide significant in EAS, but the index SNP identified in the EAS meta-analysis is in the nearby gene *EIF5A2*, encoding a protein related to translation elongation. While this could potentially be the causal gene, I believe *SLC2A2* to be a stronger candidate for having a real effect due to its role in glucose transport and therefore energy metabolism, a process known to be coupled to urate levels. If this is the case, this locus is a clear example of the potential benefits of using the MR-MEGA approach over meta-analysing separately by self-reported ancestry. The greater power and ethnic diversity has facilitated detection of a promising gene that the transethnic meta-analysis failed to detect due to the lack of a signal in Europeans.

4.4.2.2.2 *ZMIZ1*

ZMIZ1 is the second-strongest MR-MEGA unique hit, with five SNPs in the locus around the index SNP rs697238, which is nominally significant in AA, EA and EAS. This SNP is reported in PhenoScanner as being associated with type II diabetes in the UK Biobank. The encoded protein ZMIZ1 is part of the protein inhibitor of activated STAT (PIAS) family, a group of proteins that bind and regulate DNA-binding transcription factors, with ZMIZ1 reported to interact with the androgen receptor[256]. It is downregulated in patients with multiple sclerosis[257], a condition associated with reduced serum urate levels[9], though it is unclear whether it is causal[258].

4.4.2.2.3 *TAPT1-AS1*

TAPT1-AS1 is a non-coding RNA. PhenoScanner reports significant associations with acute renal failure in the UK Biobank, as well as a variety of 'cause of death' phenotypes related to gastrointestinal haemorrhage.

4.4.2.2.4 *ALDH5A1*

ALDH5A1 encodes an aldehyde dehydrogenase that catalyses the degradation of the neurotransmitter gamma-aminobutyric acid (GABA)[259]. GABA has multiple roles and is expressed highly in the insulin-producing β-cells of the pancreas, where it can

promote β-cell survival. PhenoScanner reports associations between SNPs in *ALDH5A1* and death due to B-cell lymphoma in the UK Biobank[140], and the index SNP rs73728140 has been associated with alkaline phosphatase levels[180] . Expression of *ALDH5A1* has been shown to be suppressed by hsa-miR-29a-3p (a highly abundant miRNA in human liver) in human hepatocytes. Interestingly, this miRNA also suppresses expression of *SLC22A7*, a novel serum urate locus identified in the transethnic meta-analysis[260].

4.4.3 Sex-stratified analysis

The sex-stratified analyses are perhaps most notable for their lack of significant results. Despite each sex having sample sizes in well excess of the Köttgen *et al. (*2013) GWAS, only four additional significant loci were detected, all in females, and all close to the threshold of significance. Two of these are uncharacterised lincRNAs. The index SNP in *INPP5A* is a hit for bone mineral density in UK Biobank, and the protein is involved in calcium release in mice – its deletion leads to cerebellar degeneration[261].*TRIM29* is an immune inhibitor that is co-opted by Epstein-Barr virus when infecting airway epithelial cells[262], and has been implicated as a tumour suppressor gene. The link to female-specific serum urate levels is unclear in either case and may reflect that the target genes in associations is not necessarily the closest gene.

The other noteworthy absence is in the difference in effect between sexes. *SLC2A9* and *ABCG2* have both been known since the discovery of their association with serum urate levels to carry variants with different effects in each sex – the index SNP from Köttgen *et al.* 2012 in *SLC2A9* has a stronger serum urate-increasing effect in women than in men, and the reverse is true for *ABCG2*[74]*.* In our results, the lead SNP for *SLC2A9*, rs3775947, has nearly twice as large a serum urate-increasing effect in females as in males (β_M = 0.1954, β_F = 0.3556). In *ABCG2*, rs74904971 has a 1.38-fold stronger effect in males than females (β_M = 0.2338, β_F = 0.1692).

However, considering the significant differences in serum urate levels between males and females, it is surprising that no additional loci are detected as genome-wide significant. The fact that the genetic correlation between males and females is not equal to 1, even when *SLC2A9* and *ABCG2* were excluded, supports the conclusion that there are still additional loci with undiscovered sex-specific effects. Several loci were below the suggestive significance threshold ($P < 10^{-5}$) in this analysis. These

may become significant when this analysis is repeated in, and likely meta-analysed with, the larger sample available in the UK Biobank. Future analyses could include age by sex interaction terms as an attempt to account for the effect of the menopause on urate levels in females.

Of the 183 primary meta-analysis index SNPs, six showed significant effect differences between sexes: *SLC2A9*, *ABCG2*, *CAPN1*, *GCKR*, *IDH2*, and *SLC22A12* ($P_{diff} < 0.05/183$). Sex-specific effects on uric acid at the *GCKR* locus were identified in Köttgen et al. 2013, but as here, not at the genome-wide significant level[74].

The rate of urate clearance in the kidney has been shown to be affected by varying levels of oestrogen[61–63], leading to the suggestion that sex hormones are the primary drivers of sex differences in serum urate levels. This may account for the small number of loci with significant variation between sexes if hormonal effects on serum urate are not affected by the genetic variants uncovered.

However, it seems likely that there would be genetic variants which modulate hormonal effect, and thus would appear differentially associated with women and men. For example, rs2244608, a SNP close to the *HNF1A* gene which is one of the variants with a suggestively-significant sex-specific effect, is a potential candidate for such a modulating effect, as its predicted function is to modify the binding of the oestrogen receptor and transcription factor ESR2[263–265]. The question remains open.

4.4.4 Serum urate GRS and Gout

The serum urate GRS score is clearly associated with gout in the independent UK Biobank cohort, with prevalence increasing steadily across risk score categories, and a 100-fold difference in risk between the highest and the lowest. 3.46% of the study population fell within the highest three GRS categories (≥ 5.87 mg/dL). Individuals in these categories had a greater than 3-fold increased risk for gout compared to the most common category. This is comparable to a modest effect size for monogenic disease (OR > 3)[266], but the prevalence of high GRS within the population is much higher than most monogenic disorders.

When split into a training and testing set, including the serum urate GRS showed significant improvements compared to age and sex alone in predicting gout status. It should be noted that this is not predicting future cases of gout, but rather current gout status. The GRS is still a smaller component of the risk compared to age and sex

together, but as GRS and age are fixed at birth, it may be possible to identify individuals who are at high risk of gout but have not yet developed it. In several years' time, once the UK Biobank phenotype databases have been updated with new cases of disease, it will be possible to assess whether any of our 'incorrect' predictions of gout status are in fact accurate predictions of future gout status.

Together, these finding highlight the clinical value of genetic risk scores for disease prevention – many GWAS loci have small effects but taken together they can confer a significant risk of disease on an individual. Because this risk is genetic, it remains constant throughout life, meaning an individual can be advised early in life if they have a genetic predisposition towards high serum urate levels, and thus an increased risk of gout. While this information must be explained carefully to ensure patients understand that a genetic risk does not mean they will inevitably develop the disease, there is evidence that it can lead to behavioural changes. The GeneRISK study in Finland provided patients with a web interface ("KardioKompassi") which summarised their CVD risk, including genetic risk, and followed their behavioural changes over time[267]. In a presentation at the European Society of Human Genetics meeting in Milan in 2018, they presented unpublished results showing that the greatest risk-reducing behavioural changes were seen in patients with high genetic risk. The serum urate GRS I have constructed could be used to provide patients with similar guidance on reducing their risk of gout. While not a fatal disease, gout causes considerable pain over prolonged periods, and could lead to improved quality of life for high-risk individuals if risk-reducing habits are formed early.

4.4.5 Genetic correlations between serum urate levels and published GWAS phenotypes

The strongest genetic correlations were observed with previous GWAS of serum urate – (these are actually >1, as LD-score regression does not provide a bounded estimation, but this is acceptable provided that estimates are not greatly in excess of 1). A strong correlation was seen with UK Biobank self-reported gout, which is also expected. Many positive correlations were seen with fat mass traits in UK Biobank, as well as obesity, BMI and hip circumference that are likely capturing the same effect – a genetic link between high BMI and high serum urate levels. Other strong positive correlations include traits related to triglycerides in HDL and LDL cholesterol, CKD and diabetes related traits including HOMA-IR, a measure of insulin resistance. This wide range of cardio-metabolic traits and diseases reflects

known observational correlates of serum urate. Strong negative correlations include many traits related to HDL-cholesterol, known to be negatively associated with serum urate, and, curiously, father's age at death. This suggests a possible link between low serum urate and longevity, which is in contrast to historical observations that long-lived species tend to have higher serum urate concentrations[268]. However, the known correlation between high serum urate and risk factors for cardiometabolic disease may explain this apparent contradiction in humans. A link between family longevity and low serum urate independent of kidney function has similarly been previously reported in Ashkenazi Jews[269].

No significant sex differences in genetic correlations were detected after multiple testing correction, likely due to the relatively small sample sizes, but nine traits were nominally significant ($P < 0.05$). Larger GWAS sample sizes would lead to smaller SEs in the calculation of r_g, which may lead to some of these suggestive signals becoming significant.

Sex-differences in many of the phenotypes make intuitive sense – for example, these results suggest that variants associated with later menopause are associated with lower serum urate levels in women but show no association in men. This could be confounded by BMI and hormonal levels both positively correlating with serum urate levels but does lead to an interesting possibility – that the shared variants between serum urate and age of menopause are not affecting serum urate levels in men. This would support the intuitive notion that there are sex-specific serum urate loci beyond those that we have identified, perhaps regulating the interaction between sex hormones and serum urate levels.

4.4.6 Limitations of annotation

The number of hits makes the downstream analysis of these results a complex prospect. The Köttgen *et al*. analysis identified eighteen novel loci, a small enough number for each to be manually investigated in detail by a team of analysts, but already a formidable task. Understanding individual roles of the 147 loci identified in this analysis is a far greater undertaking, and manual annotation and literature searching becomes an unrealistic prospect within the timescale of a single publication, or even a PhD thesis. Instead, large GWAS projects are increasingly turning to bioinformatics tools to interpret the highly polygenic associations. Tools such as DEPICT and FUMA aim to integrate some of the vast quantity of publicly-

available information to link GWAS hits to known pathways, expression data and work on model organisms. They have proven to be versatile and useful methods, but there are problems with this approach.

The first is that these frameworks are limited by their methodology – any approach to reducing the dimensionality of a dataset will inevitably result in the loss of information, and the mathematical and statistical methods used to do this will favour certain kinds of information over others, in a way that may not always be obvious to a user.

The second is that an information aggregator can only ever be as good as the information it aggregates. Even during the process of writing this thesis I have found flawed information in the PhenoScanner database, where the GWAS results from the OLINK-IMPROVE study are clearly incorrect. Nonetheless these results have been integrated into the PhenoScanner database, and incorrect conclusions may have been drawn from a pQTL that simply did not exist. Databases have to be kept up to date, and tools have to be maintained. One can use multiple tools to compensate for each individual approach's weaknesses, but the user quickly arrives back at the problem of having too much data to process manually.

4.4.7 The 'omnigenic' hypothesis

GWAS signals are often spread across the genome, and the implicated genes are rarely obviously connected to the disease or trait in question. One interpretation of the complexity of GWAS results that has generated a lot of discussion in recent months is the 'omnigenic' hypothesis put forward by Boyle *et al.* in 2017[270]. They propose that instead of trying to understand each hit in isolation, gene regulatory networks can be used to identify 'core' genes, which have a direct effect on the complex trait. Because all the genes expressed within the relevant cell-type will have some small impact on the function of these core genes, and so explain the bulk of the heritability of the trait without having a direct impact on the trait itself. Under this framework, attempting to understand individual GWAS hits for non-core genes in isolation is a fruitless exercise, at least as far as drug discovery is concerned.

The authors suggest that a more useful approach is to use Whole Exome Sequencing to search for rare variants of large effect which may tag core genes, and to focus on the construction of cell-type specific gene-regulatory networks. Several recent publications have adopted the omnigenic hypothesis to explain their findings[271,272], but the paradigm has not gone unchallenged. In particular, Wray *et al.*[273] believe it does

not reflect our current knowledge of the polygenicity of complex traits and diseases, although they agree that the prioritisation of gene-regulatory networks is valuable.

Personally, I do not find the omnigenic hypothesis to be particularly compelling – partly because I believe the term is redundant when 'polygenic' already covers most of the key concepts, but primarily because the theory hinges on the assumption that a small number of core genes are driving variation in a trait, which, while compelling and undeniably neat, I do not believe we can state to be true.

Interpreting uric acid under the omnigenic model might initially seem quite intuitive. As it is linked to so many metabolic processes, one might expect to see transcription factors that control the metabolic rate in the cell and the production of waste products as the core genes at the centre of a wide network. We have in fact identified some of these in our meta-analyses, for example *HNF1A* and *HNF4A*, major transcription factors in the liver and the kidney proximal tubules – the relevant tissues for urate – and they do indeed regulate a large number of other genes, many of which are unrelated to uric acid levels. However, while these genes definitely have an effect on serum urate levels, to consider these the core genes would be to dismiss the urate transporters as less important peripheral genes. This would miss their importance as key targets for pharmaceutical interventions.

Consequently, some have suggested the omnigenic model is best viewed as a gradient from 'core' genes to peripheral genes, with varying influence on the traits – which, in my view, is indistinguishable from the polygenic model we are already using.

In the case of serum urate, a more useful concept to understand the results of our meta-analysis may be that of pleiotropy – the same variant or gene having an effect on multiple different traits. Many of the loci identified in our results are highly pleiotropic, which may be indicative of co-regulation of related processes between serum urate and other traits. Understanding the mechanisms driving this pleiotropy may lead to new possibilities for urate-altering pharmaceutical interventions that avoid disrupting other processes in the cell.

4.4.8 The future of GWAS

The UK Biobank has proven massively disruptive to the status quo in the GWAS field. In part because any individual can now run a large GWAS with minimal investment of time, money or manpower. More significantly, however, several groups have

automated pipelines for GWAS of almost all the phenotypes included in the UK Biobank catalogue and made these results available online in tools such as GeneAtlas[274], the Global Biobank Engine[275] and PhenoScanner[140]. In many cases, it is now unnecessary to run a straightforward GWAS on a trait if it has already been included in one of these databases.

Consortia have monopolised the GWAS field for several years now, but almost overnight, their era may be coming to an end. A regrettable side effect of the wide availability of UK Biobank data may be a reduction of willingness to collaborate with others – instead of almost every scientist in the GWAS field working together by necessity in one consortium, each individual can now operate alone without sacrificing access to data. Instead of collaboration, labs must compete to be the first to publish an analysis. And the UK Biobank is merely the first 'supercohort' – others are close behind, including the ambitious Million Veteran Program, a self-proclaimed 'mega-biobank' that aims to recruit one million US veterans, linking genotypes to rich data from healthcare providers[276]. Soon even the UK Biobank may be dwarfed.

This is not to say that GWAS consortia will disappear. These groups represent much of the talent within the field, and each one brings together the collective knowledge of experts within their specialist area. A GWAS is just a statistical technique. Knowing how to prepare the phenotypes – who to exclude, how to transform the data, what covariates will be relevant – is more complex, and beyond the simple analyses performed by the UK Biobank GWAS databases currently available. The future of the field may lie in more sophisticated analyses that integrate data that is not publicly available, such as the kidney eQTL mapping and transethnic analyses included in our paper.

The field is undoubtedly changing, but the end of GWAS has been predicted before. There has been a pattern in recent years of running bigger and bigger GWAS, and while this has led to the identification of increasingly large numbers of variants, it is unlikely to be sufficient to carry a project. Consortia will have to identify their unique strengths if they wish to remain competitive in the era of biobanks.

Chapter 5 Gout risk in the UK Biobank

5.1 Background

5.1.1 The UK Biobank cohort

The UK Biobank is a recent publicly-available prospective cohort with rich phenotypic data on over 500,000 participants of UK origin[75,234]. Genotypic data has been released for all 500,000 participants on a custom panel combining the HRC imputation panel with structural variants from the 1000 Genomes Project.

The cohort is noteworthy for two aspects: its size, and its open availability. For a nominal access fee, and subject to approval of a proposed project, any researcher in the world can gain access to a population cohort of half a million individuals with rich phenotypic data. This allow statistical power previously only available to consortia meta-analysing dozens of smaller cohorts, which carries a significant administrative burden and has greater risk of heterogeneity due to varying techniques between groups. Any follow up analyses must either be performed only in whatever small cohorts the lead analysts have direct access to, or else must be sent back to all collaborators, a process which can take as long as the primary analysis to complete. In contrast, the UK Biobank is a single cohort, with consistent techniques applied for the measurement of all phenotypes. Complete data on all participants is available to the analyst – and often there is only one – making more complex downstream analyses such as conditional testing to identify distinct signals much more practical, and considerably faster.

Unfortunately, the unprecedented scale of the UK Biobank project means that it is something of a trailblazer, inevitably leading to unforeseen delays and setbacks. Most regrettable of these is the delay in the release of the blood biochemistry data. Originally scheduled for release in mid-2017, this has been pushed back several times and at the time of writing remains unreleased with no anticipated schedule. The delay has been due to quality control problems that stem from scaling assays to half a million individuals, and the data is as yet unsuitable for analysis, though UK Biobank reports that it will still be released at some point.

This data unfortunately includes measurements of serum urate, which formed a component of the original proposal for my PhD. I was intending to take advantage of

the size of the cohort to perform a series of stratified GWAS – dividing the cohort based on phenotypes known to be related to uric acid. This was to begin with a sex and BMI-stratified analysis, that would develop a smaller initial analysis published by our group in 2015[69], and would likely have also included stratification by alcohol consumption. I also considered imputing phenotypes into the UK Biobank for the proteins identified in Chapter 2 using the genetic variants identified in meta-analysis (Section 2.3.5).

At the time of writing, the UK Biobank biochemistry data remains unreleased. The serum urate measurements would have been a valuable resource both for supporting the analyses described in this thesis and for performing standalone analyses, but the cohort has still been useful without it. Gout status is available from self-reported data, and additionally hospital admission records include ICD10 codes which can be used to identify gout cases (code M10). This information was used to test association between the CKDGen serum urate GRS and gout (Section 4.2.5). UK Biobank genetic data was also used as a reference population for estimating LD structure to allow approximate conditional analysis in Europeans.

5.1.2 Gout case/hypernormal control analysis

Most people with hyperuricaemia do not develop gout. Consequently, there is considerable interest in identifying commonalities between high risk individuals who do not develop the disease that distinguish them from those who do. One approach to identify genetic factors affecting gout propensity is to use hyperuricaemic controls in a case-control GWAS. If both cases and controls are hyperuricaemic, differences between them may reflect other aspects of gout such as monosodium urate crystal deposition and inflammatory response to crystals. A project using this approach is currently being coordinated by Professor Tony Merriman (University of Otago).

Serum urate is not the sole factor affecting gout risk, however, with diet as a major factor and sex, age and ethnicity all contributing. In a study on the National Health and Nutrition Examination Survey 2007-2008 (a survey assessing the health a nutrition of people in the USA), of the 5,707 men and women over 20 with gout, 74% had hypertension, 71% CKD stage two or greater, 53% were obese, 26% had diabetes, 24% nephrolithiasis (kidney stones), 14% had suffered myocardial infarction, 11% had heart failure, and 10% had suffered a stroke. For all comorbidities, the proportions were significantly higher than in non-gouts[277].

While the biochemistry data remains unavailable in the UK Biobank, many phenotypes are available that are (or co-occur with) comorbidities of gout. By using a linear combination of these phenotypes, I have constructed a predictive model that generates a gout risk score for an individual. In contrast to the serum urate GRS in Section 4.3.6, which assesses the utility of SNPs associated with serum urate for predicting gout, this phenotypic gout risk score is intended to identify high-risk non-cases – so called 'hypernormal' controls – for use in a case-control GWAS. This analysis should detect genetic variants that confer increased resistance or susceptibility to gout independently of the environmental risks included in the model and has the additional benefit of being possible before the biochemistry data is released.

This project is still under development in collaboration with Professor Tony Merriman and Dr Tanya Major (University of Otago). I have used the Generation Scotland cohort to construct a gout prediction model and applied this model in the UK Biobank to identify a list of individuals who could be used as high-risk controls. This information has been sent to our collaborators in Otago, who will run the GWAS. Final results are not yet available, but the development of the prediction model and its findings are presented below.

5.2 Methods

5.2.1 Generation Scotland

The Generation Scotland: Scottish Family Health Study is described in Section 3.2.1.4. For the purposes of this analysis, only unrelated individuals were used. This was assessed by removing one of each pair of individuals with a kinship coefficient of greater than 0.015625 (1/64, corresponding to the average shared kinship between second-degree cousins) in an identity-by-state kinship matrix calculated using the 'ibs' function in the *GenABEL* R package[132]

Gout status was obtained from a combination of self-report, prescription of serum urate-reducing medication from NHS prescription data (allopurinol, febuxostat, benzbromarone or probenecid) and SMR01 hospital admissions data (any instance of ICD10 code M10, corresponding to gout).

5.2.2 Phenotypes

The initial set of phenotypes to be tested was based on those available in both Generation Scotland and UK Biobank – this unfortunately ruled out eGFR, as creatinine is not yet publicly released in the latter.

The initial input phenotypes were age, sex, BMI, height, weight, waist-to-hip ratio, body fat percentage, systolic and diastolic blood pressure (SBP and DBP respectively) (averaged across two readings and corrected for blood-pressure reducing medication by adding 15 and 10 mm Hg to SBP and DBP, respectively, for individuals reported to be taking BP-lowering medication, as per the protocols used by the International Consortium for Blood Pressure[278]), heart rate (averaged across two measurements), hypertension calculated from blood pressure measurements (SBP ≥140 mm Hg or DBP ≥90 mm Hg), self-reported high blood pressure and units of alcohol consumed per week.

After preliminary testing, to reduce the number of combinations required for testing, this set of phenotypes was reduced to age, sex, BMI, waist-to-hip ratio, body fat percentage, self-reported high blood pressure, hypertension calculated from corrected blood pressure measurements and units of alcohol consumed per week. Only individuals with no missing phenotypes were retained in the analysis.

5.2.3 Linear regression

The R package 'bestglm'[279] was used to regress gout status on all possible combinations of covariates in Generation Scotland using logistic regression models. The best model was selected based on lowest value of both the Akaike information criterion (AIC). The AIC is an estimator of the quality of a statistical model of a set of data and is intended to balance goodness-of-fit against overfitting. Bestglm calculates the AIC automatically and uses this to select the best model.

5.2.4 Gout risk score in UK Biobank

The best model was then used to generate a gout risk score from phenotype data in the UK Biobank using the 'predict' function from the package *stats*. Accuracy of prediction was assessed with the AUC of a ROC curve using the model 'gout ~ gout risk score'. This analysis was performed under UK Biobank projects 8304 and 12611.

Gout risk scores were given to the Merriman group to allow selection of the best threshold for classifying an individual as a hypernormal control, depending on the number required. Proposed thresholds included the median, mean and 3rd quartile of the risk score across the whole cohort.

5.2.5 Comparison to serum urate risk score

Establishing whether the gout risk score merely identifies hyperuricaemic individuals would ideally be done by comparing serum urate levels to the score. In the absence of this data, I constructed a serum urate genetic risk score (GRS) as a proxy measure of serum urate levels in UK Biobank individuals. This score was constructed as described in 4.2.5, but as the results of the CKDGen meta-analysis are unpublished and this project is an independent effort, I instead used SNPs and effect sizes reported in Köttgen *et al.* (2013)[74].

5.3 Results

5.3.1 Generation Scotland phenotype summaries

Phenotype preparation in Generation Scotland lead to a dataset containing 147 unrelated gout cases and 6223 controls. Summaries of quantitative phenotypes are given in **Table 31**. The analysis dataset contained 2,772 males and 3,598 females, with 920 cases of self-reported high-blood pressure and 2,453 cases of hypertension calculated from blood pressure measurements. Body fat percentage included some unlikely extremely low values, possibly due to inaccurate readings from the bioimpedance machine. However, most values are biologically feasibly, so in the absence of a strong justification for a threshold to remove the low values, I have chosen to retain them.

Table 31 – Summary of quantitative phenotypes tested for gout risk model in Generation Scotland.

	Age	BMI	WHR	Body Fat %	Alcohol (units/week)
Min	18	16.02	0.41	1	0
Median	51	26.03	0.86	29.8	8
Mean	49.35	26.78	0.87	30.06	10.79
Max	92	56.6	2.23	57.6	215
SD	13.72	5.03	0.09	9.39	12.50

5.3.2 Gout risk model

The best logistic regression model contained terms for age, sex, BMI, self-reported high blood pressure and alcohol consumption. The coefficients for this model are given in **Table 32**.

Table 32 – Coefficients included in the gout risk score model

Coefficients:	Estimate	Std. Error	P-value
(Intercept)	-11.8634	0.7550	$< 2 \times 10^{-16}$
Age	0.0601	0.0084	9.99×10^{-13}
Sex (Male)	1.6819	0.2275	1.44×10^{-13}
BMI	0.1190	0.0154	1.30×10^{-14}
S.R. High Blood Pressure	0.6439	0.1876	6.0×10^{-4}
Alcohol (units/week)	0.0139	0.0051	6.28×10^{-3}

5.3.3 Correlation with serum urate in Generation Scotland

Gout risk scores were calculated for Generation Scotland and compared to serum urate measurements obtained from EHRs (described in Section 3.2.1.4). Overlap between individuals with serum urate measurements and those with complete phenotypes for generating gout risk scores was low at only 126 individuals. No significant correlation was observed between serum urate levels and gout risk score (Spearman's r = 0.0568, p = 0.528). Correlations were stronger when separated by sex, with Spearman's r = 0.0933 for females and 0.153 for males (**Figure 40**), but the correlation was still not significant (p = 0.0436 and 0.270 respectively).

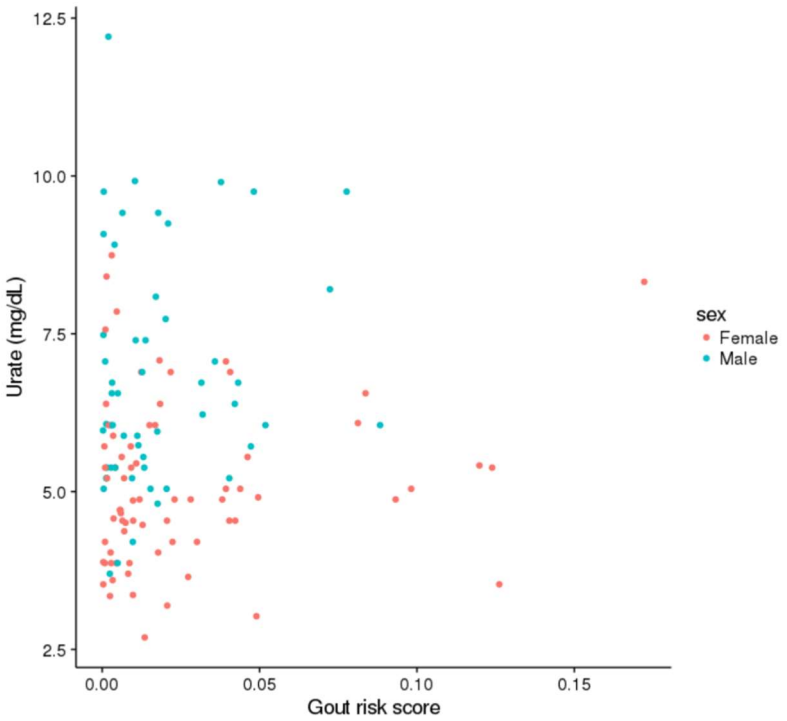


Figure 40 - Scatterplot of gout phenotypic risk score against serum urate level in Generation Scotland.

5.3.4 UK Biobank phenotypes

Gout status was provided by the Merriman group (UK BB project 12611), and was ascertained from self-report, hospital admissions and prescription of serum urate-lowering medication. Individuals with no information available on gout status were removed, leaving 105,421 individuals with definite gout status. After filtering to retain only individuals with complete phenotypes for gout prediction, 73,015 remained. The majority of individuals removed at this stage were missing alcohol consumption data. Summaries of the quantitative phenotypes used to generate gout risk scores are given in **Table 33**. The analysis dataset comprised 33310 females to 39705 males, of whom 19,245 self-reported high blood pressure.

Compared to Generation Scotland, average age is similar although the range of ages represented is narrower. BMI is comparable, but alcohol consumption appears to be slightly higher in the UK Biobank. Both cohorts have a few extremely high-consumption individuals.

Table 33 – Summary of quantitative phenotypes used to generate gout risk score in UK Biobank.

	Age	BMI	Alcohol (units/week)
Min	40.00	14.67	0.00
Median	58.00	26.50	16.80
Mean	56.88	27.04	22.72
Max	73.00	61.00	239.40
SD	7.84	4.33	19.11

5.3.5 UK Biobank risk scores

Risk scores were generated using the model specified in **Table 32**. The distribution of risk scores is displayed in **Figure 41**, and it can be clearly observed that while gout cases have higher risk scores on average than controls, there are still large numbers of controls with high risk scores.

The model was also used to predict gout status as a means of testing the model in an independent population sample. The ROC curve comparing predicted to actual gout status is given in **Figure 42**, with an AUC of 0.8441, suggesting the model predicts gout status well.

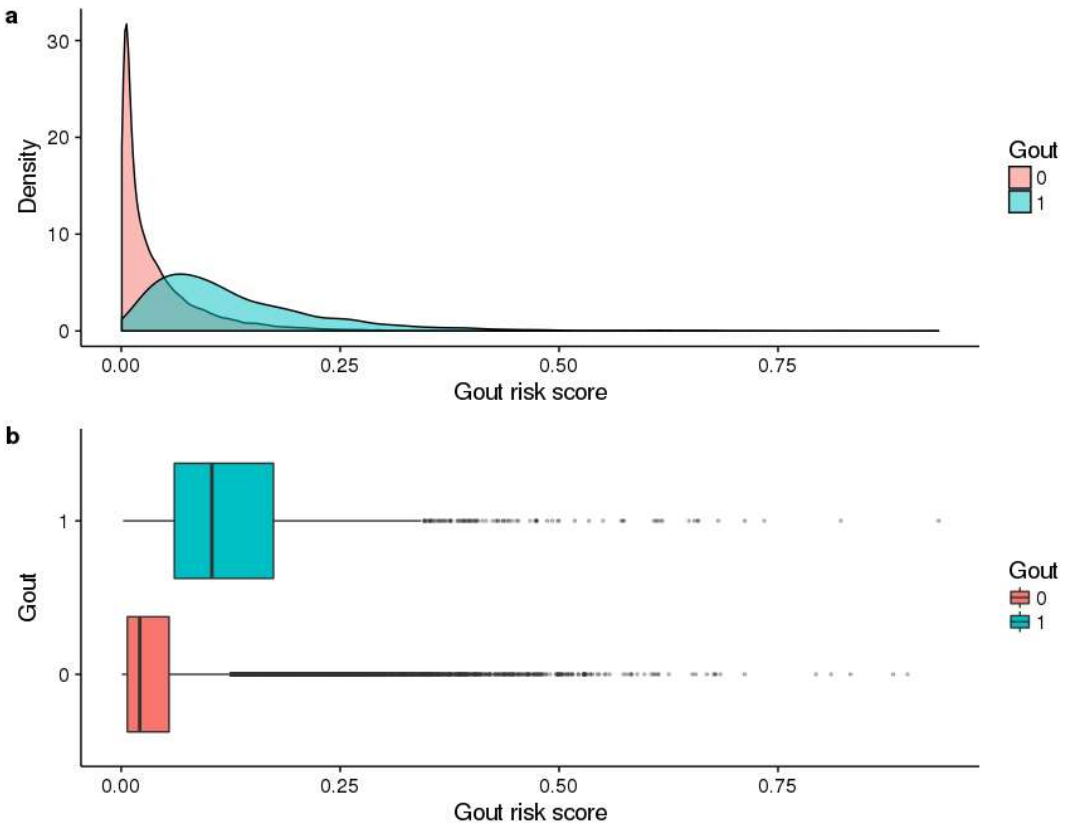


Figure 41 - Gout risk score distribution in UK Biobank.

a Kernel density plot of gout risk score split by gout case (1) vs control (0) status.

b Boxplot of gout risk score by gout case/control status.

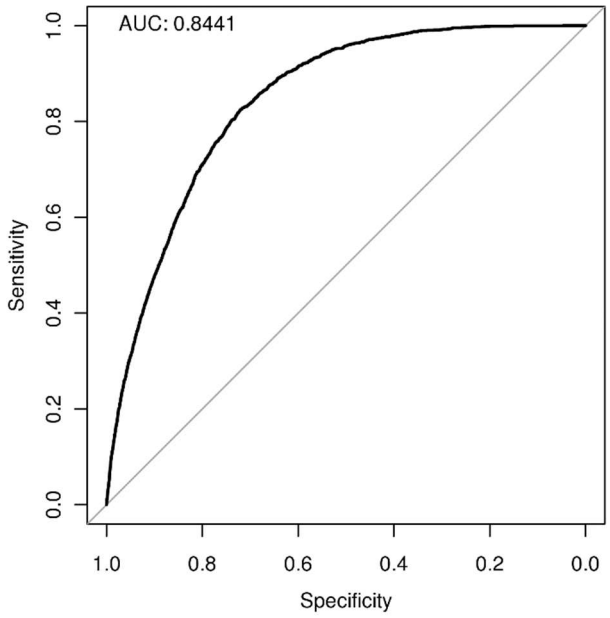


Figure 42 - ROC curve for gout status prediction in UK Biobank from gout risk score.

5.3.6 Relationship to serum urate GRS

The serum urate GRS was completely uncorrelated with the gout risk score across the whole sample (Spearman's r = 0.00173, **Figure 43**). The correlation was stronger in gout cases only, but still relatively weak (Spearman's r = 0.102).

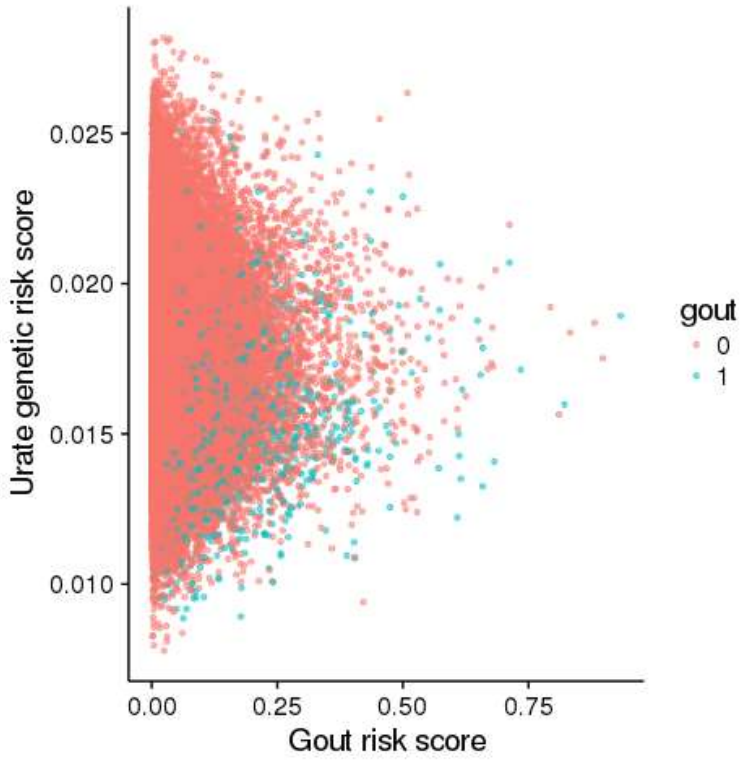


Figure 43 - Scatterplot of gout risk score against serum urate genetic risk score in the UK Biobank.

5.3.7 Defining hypernormals

One possible example for defining hypernormality would be to take a threshold of the 3rd quartile of the gout risk score (>0.0577). This give a ratio of 1,961 gout cases to 16,761 hypernormal controls. One could also apply the risk score cut-off to gout cases as well, to exclude gout cases where the cause is not reflected in the comorbidities used to generate the score, and ensure the risks are comparable between groups, leaving 1,493 cases.

A more stringent threshold could be the 90th percentile (>0.114, 879 cases, 6423 controls) or a more biologically meaningful one the median risk score value of gout cases (>0.103, 980 cases, 7572 controls). Exploratory genetic analyses using these are to be made by the Merriman group.

5.4 Conclusions

The use of 'hypernormal controls' has been criticised in the field of psychiatric genetics, where it has been claimed to introduce bias[280]. In this example, a hypernormal control is a healthy individual with no history of psychiatric disorders at all, while the case can have any number of disorders as long as this includes the one of interest. Thus, it is impossible to distinguish between an association driven by the trait of interest and an association driven by another disorder – by selecting controls with no disorders at all, one increases the risk that many phenotypes differ between cases and controls. This concern has been noted more generally, as a caution that while a hypernormal control study can be more powerful, one must be aware of possible bias when selecting controls[281].

To avoid the problem of introducing bias, our analysis should apply the same score threshold to both cases and controls. Although this reduces the sample size, the risk of detecting loci associated with the phenotypes used to generate the score is also reduced. In this sense, our analysis will be closer to a risk-matched control analysis and caution should still applied when interpreting the results.

When the Merriman group run this exploratory GWAS, it will be interesting to see whether the results identify loci not seen in a parallel case-hyperuricaemic-control analysis. While the overlap between individuals with complete phenotypes for calculating gout risk scores and individuals with serum urate measurements is small in Generation Scotland, there appears to be no correlation in this cohort. In UK Biobank, the comparison of the gout risk score with the serum urate genetic risk score suggests that the gout score is capturing an element of risk that is distinct from an individual's genetic predisposition to hyperuricaemia. The GRS will of course not completely capture serum urate levels as environmental factors will also affect this, but as observed in Section 4.3.7.2, a serum urate GRS provided additional predictive power for gout over just age and sex.

In the interim, the results of the gout phenotypic risk score are presented here. By using only non-genetic, non-urate risk factors, I have been able to construct a model that predicts gout status with an AUC of 0.84, comparable to the performance of the age + sex + GRS model used in Section 4.3.7.2. It is my belief that the 'misclassified' individuals, those predicted to have gout but who do not, may represent the best set of individuals to uncover novel genetic variants protective of gout.

Chapter 6 Conclusions and summary

In this thesis, I have combined two complementary approaches to improve and extend our understanding of uric acid and the role it plays in health and disease – its relationship with other phenotypes, specifically endophenotypes related to disease, and its relationship with genotype.

In Chapter 2, eleven protein biomarkers for cardiovascular disease were identified as being significantly correlated with serum urate levels independently of all the remaining 266 proteins tested, as well as age, sex, BMI, eGFR and alcohol consumption. Some of these expand on previously reported associations, including FGF-21 and FGF-23, demonstrating that the correlations are not explained by any of the wide range of covariates included in my models. In other cases, implicit associations have been reported, such as that between the uricosuric drug benzbromarone and FABP4, but no explicit correlation with serum urate has been identified. In many cases, my results represent, to the best of my knowledge, the first time that these associations have been identified.

The protein biomarkers were selected for their association with cardiovascular disease or inflammation – and intriguingly, despite the pro-inflammatory role of uric acid, no proteins from the INF panel were identified in any of my analyses. Of course, CCL3 and IL-1RA are both proteins that play a role in inflammation, but both were on the CVD panels. This perhaps hints that the inflammatory role of serum urate is sufficiently closely associated with cardiovascular disease that the serum urate-related inflammatory proteins are biomarkers for CVD.

I have also investigated associations between serum urate and levels of specific lipid species, identifying a negative correlation with PC 38:6, a lipid associated with progression of Alzheimer's disease and Huntington's disease. Several other lipid associations were also identified.

Despite the targeted set of proteins initially queried, a diverse range of biological processes are implicated in the results, including phosphate metabolism and bone development, glucose metabolism, adipocyte function, blood pressure regulation. In many cases the link with serum urate is known, but the mechanism is only partially understood. More targeted analyses, for example, in the case of FGF-21 and *GCKR*, could establish whether the connection with serum urate is driven by pleiotropy or whether the relationship is causal. The release of the SCALLOP consortium's GWAS

results for CVD-II, CVD-III and INF proteins will facilitate greatly expanded querying of associated pQTLs.

Chapter 3 began to explore the relationship between variation in the genome and serum levels of uric acid in our in-house cohorts, a sample of over 10,000. Though larger GWAS have been run in the past, I identified a number of loci that were unreported at the time of analysis, in part due to the increased density and accuracy of the HRC imputation panel, which allows SNPs closer to the true causal variant to be tested for association with the phenotype.

In addition to this, this chapter showcases the capacity of electronic health record linkage in Generation Scotland to provide valid phenotypic data for use in analyses. Clinical measurements of uric acid from NHS blood tests for 2,077 individuals were used in a GWAS of serum urate which identified the known strong *SLC2A9* signal and confirming its validity, and also identified novel signals in three additional genes. These results were published as part of Nagy *et al.* (2017)[206], and the EHR linkage is described more fully in Kerr et al. (2017)[282]. We also identified an unusual signal in the ORCADES cohort on chromosome 11, which we hypothesise may be caused by a pericentric inversion containing the gene *SLC22A12*, encoding the known urate transporter URAT1. Though we cannot easily test this hypothesis, we will soon have access to whole genome sequencing data which might identify rare coding or structural variants in URAT1 in individuals who appear to drive this signal.

In Chapter 4, I have expanded on this exploration of genotypic associations with serum urate, reporting my contributions to the large meta-analysis of over 450,000 individuals run by the CKDGen consortium, which at the time of writing is about to be submitted for publication. This major collaboration is the work of many authors, but I have played a central role in the meta-analysis itself, as well as running many of the secondary analyses. We have identified 183 loci genome-wide, with 147 considered novel – a five-fold increase in the number of loci known to influence serum urate levels, and a major resource that will be made available to the broader scientific community.

When the UK Biobank biochemistry data is released, it will likely not be long before GWAS of these new phenotypes are run, including urate – this will mean a sample size of up to half a million individuals, almost certainly exceeding the size we have achieved with our meta-analysis, and requiring much less analyst time. There is an

inevitable question which must be asked: was our analysis worth it? I believe that it was, for three reasons.

Firstly, previous rounds of data release from the UK biobank have shown an early 'gold rush', where groups race to be the first to publish GWAS results for new phenotypes. For example, a pre-publication manuscript of an analysis of the UK Biobank depression phenotypes was released on bioRχiv (www.biorxiv.org, a pre-print server for biological sciences, where researchers can upload papers that have not yet been accepted for publication, both to share their results as early as possible, and to establish precedent and avoid competition with other groups) within a single week of the data being published[283]. In this case, the scope was limited to a small set of phenotypes and the analysis thorough, but in others, automated pipelines have been used to run a GWAS on every phenotype available, and the results released as a 'resource' without further comment[140,274,275]. While these easily-queried databases are undeniably a valuable resource – I have made use of them myself several times in this thesis – they provide no interpretation of the results. All the user gets is an association and a P-value for a SNP with a trait. In contrast, a paper such as ours develops the analysis further, using the combined experience of a large analyst group to interpret results, and performing downstream fine-mapping and annotation as well as expert knowledge of phenotypes studied.

Secondly, while the UK Biobank is a powerful resource, it represents a primarily European-ancestry sample, with 94.6% of participants self-reporting their ethnicity as "White British", a slightly higher proportion than the UK as a whole (91.3%)[284]. While representative of its country of origin, it covers only a limited part of the spectrum of genetic diversity worldwide – a criticism that can be aimed at the GWAS field in general. The political and scientific factors driving this have been debated[285–287] but it is clear that there is a disadvantage to missing this diversity. For example, the PAGE study looked at variants identified in Europeans as being associated with BMI, lipid levels and type 2 diabetes in five non-European populations, and found that 25% of these had a different effect in at least of one population[288]. Differences in LD and haplotype frequency across populations can also be useful tools for fine-mapping. While our study is still predominantly European, it contains a significant non-European component, which we have been able to use to identify loci that would otherwise have been missed. Using my transethnic meta-regression analysis implemented in MR-MEGA, I was even able to use data from multiple ethnicities to identify an association

with the *SLC2A2* locus, a region that was identified in East Asians separately, but with a lead SNP in the less promising candidate *EIF5A2*.

Finally, the fact we ran our GWAS on a separate sample to the UK Biobank allows it to be used as a resource for secondary analyses. Because the effects for the SNPs used to create the risk score were established in a separate population, we avoid having to choose between dividing our sample into a training and testing set, sacrificing power, or running the GRS regression in the same data set used to build the score, risking overfitting of the model, and failing to prove its generalisability.

Additionally, when the UK Biobank data is released, as the overlap between our sample and the UK Biobank is minimal (though non-zero, as there are at the least a small number of individuals represented in both UK Biobank and Generation Scotland), it will be possible to meta-analyse the GWAS results from our project together with the new data. This combined sample will approach a million individuals and will likely bring a whole new wave of associated loci, which might help to address the question of whether urate follows the 'omnigenic' structure.

This analysis has also fed back into the results from Chapter 2, with the meta-analysis used to look up SNPs associated with serum levels of the protein biomarkers significantly correlated with uric acid. A promising association was at the *GCKR* locus for FGF-21, but no colocalisation of signals was detected. However, since this locus was only suggestively associated with FGF-21, it will be informative to repeat using the results from the larger GWAS currently being run by SCALLOP.

The field of statistical genetics is constantly evolving, with new tools and methods constantly being developed and refined. Even in the time since completing analysis on this thesis, a new method has been published by O'Connor and Price for distinguishing in genetic correlations genetic pleiotropy from causation[289], avoiding the confounding that creates problems in MR from shared aetiology between traits. This method that could be applied to the tests for genetic correlation between protein biomarkers and serum urate, once the SCALLOP consortium has published GWAS of sufficient size to apply the method. Adrienne Tin is currently running this analysis on the CKDGen urate results, to supplement the publication. The method currently lacks the straightforward and user-friendly implementation of LDSC or LD-hub, the other Price group methods, making it difficult to implement at this late stage, but it serves to illustrate that this field is under constant development. Our published results will hopefully be integrated into many projects to come.

Chapter 5 is a review of the progress on a collaborative project with the Merriman group at the University of Otago. This project began as an exploration of how to make use of the UK Biobank cohort before the release of the biochemistry data. While the data release has been pushed back multiple times over the last three years, with no release date yet confirmed, this project has developed into a thriving collaboration. The use of controls with a high phenotypic risk for gout should complement their ongoing analysis using hyperuricaemic control, as our preliminary work appears to show that these capture different elements of gout risk. That gout status can already be predicted quite accurately in the UK Biobank is also of interest. The comorbidities of gout are well known, but I to my knowledge no one has published their predictive ability in a general healthy population – although I have found an abstract from the 2013 American College of Rheumatology meeting[290] that predicted incident gout in hyperuricaemics. My model was able to predict gout status more accurately, but as I am predicting prevalent rather than incident gout, a true comparison will not be possible until more data is available on incident cases from the cohort. I look forward to seeing the development of this analysis.

These separate chapters explore in detail different facets of two central questions – how is urate metabolism and homeostasis regulated, and how does its dysregulation lead to disease? Serum urate levels are a convenient proxy measure for a complex integration of physiological processes, including metabolism of urate, renal excretion and reabsorption, elimination in the intestine and dietary intake of purines. I have identified a large number of new regions of the genome associated with variation in serum urate levels, both independently and as part of a large GWAS consortium. Though we have interpreted these results as best we can, and our results support a strong role for the kidney in regulation of serum urate levels, it will likely take many years before the roles of the loci we have detected are fully understood. I have also identified multiple protein biomarkers for cardiovascular disease that appear to be linked to changes in serum urate levels, which, with further investigation of pleiotropy and causality, may help elucidate the mechanisms that connect hyperuricaemia with CVD.

On the question of further work, in particular I would like to further understand the link between serum urate levels and the proteins I identified, starting by testing whether they persist when a different platform is used to quantify them. The INTERVAL cohort has offered to apply my technique to their SOMAscan data, which measures an order

of magnitude more proteins than the Olink data we have available. If associations were found to persist on a different platform and with the levels of a much wider range of proteins taken into account, their validity would be strongly confirmed. The release of more detailed pQTL data from SCALLOP will also testing of causality with MR and other techniques. And finally, when the UK Biobank serum urate data is finally released, genetic associations for these proteins could be used to impute their levels in this very large cohort, allowing stratification and interaction analyses with uric acid levels – an early aim for this project, before the delay made this unfeasible. As with many endeavours in science, I conclude this project with as many new questions as I have answers.

In summary, this thesis has contributed to the study of uric acid homeostasis and its related disease through exploration of its relationship with both genotype and with endophenotypes related with disease.

X. References

1. Fathallah-Shaykh, S. A. & Cramer, M. T. Uric acid and the kidney. *Pediatr Nephrol* **29**, 999–1008 (2014).
2. Nuki, G. & Simkin, P. A. A concise history of gout and hyperuricemia and their treatment. *Arthritis Res Ther* **8**, S1 (2006).
3. Macnalty, A. S. A Short History of the Gout and the Rheumatic Diseases. *Med Hist* **8**, 394–395 (1964).
4. Scheele, C. W. Examen Chemicum Calculi Urinari [A chemical examination of kidney stones]. *Opuscula* **2**, 73 (1776).
5. Mccarty, D. J. & Hollander, J. L. Identification of urate crystals in gouty synovial fluid. *Ann. Intern. Med.* **54**, 452–460 (1961).
6. Johnson, R. J. *et al.* Sugar, uric acid, and the etiology of diabetes and obesity. *Diabetes* **62**, 3307–3315 (2013).
7. Ndrepepa, G. Uric acid and cardiovascular disease. *Clin. Chim. Acta* **484**, 150–163 (2018).
8. Indraratna, P. L., Williams, K. M., Graham, G. G. & Day, R. O. Hyperuricemia, Cardiovascular Disease, and the Metabolic Syndrome. *The Journal of Rheumatology* **36**, 2842–2843 (2009).
9. Spitsin, S. & Koprowski, H. Role of uric acid in multiple sclerosis. *Curr. Top. Microbiol. Immunol.* **318**, 325–342 (2008).
10. de Lau, L. M. L., Koudstaal, P. J., Hofman, A. & Breteler, M. M. B. Serum uric acid levels and the risk of Parkinson disease. *Ann Neurol.* **58**, 797–800 (2005).
11. Pulido, R., Jiménez-Escrig, A., Orensanz, L., Saura-Calixto, F. & Jiménez-Escrig, A. Study of plasma antioxidant status in Alzheimer's disease. *Eur. J. Neurol.* **12**, 531–535 (2005).
12. Tana, C., Ticinesi, A., Prati, B., Nouvenne, A. & Meschi, T. Uric Acid and Cognitive Function in Older Individuals. *Nutrients* **10**, (2018).
13. Ahmed, M., Taylor, W., Smith, P. R. & Becker, M. A. Accelerated Transcription of PRPS1 in X-linked Overactivity of Normal Human Phosphoribosylpyrophosphate Synthetase. *J. Biol. Chem.* **274**, 7482–7488 (1999).
14. Petrie, J. L. *et al.* The rate of production of uric acid by hepatocytes is a sensitive index of compromised cell ATP homeostasis. *American Journal of Physiology-Endocrinology and Metabolism* **305**, E1255–E1265 (2013).
15. Kushiyama, A. *et al.* Role of Uric Acid Metabolism-Related Inflammation in the Pathogenesis of Metabolic Syndrome Components Such as Atherosclerosis and Nonalcoholic Steatohepatitis. *Mediators of Inflammation* (2016). doi:10.1155/2016/8603164
16. Ramazzina, I., Folli, C., Secchi, A., Berni, R. & Percudani, R. Completing the uric acid degradation pathway through phylogenetic comparison of whole genomes. *Nature Chemical Biology* **2**, 144–148 (2006).
17. Benedict, S. R. The Harvey Lectures. *Journal of Laboratory and Clinical Medicine.* **1**, 346 (1916).
18. Kuster, G., Shorter, R. G., Dawson, B. & Hallenbeck, G. A. Uric Acid Metabolism in Dalmatians and Other Dogs: Role of the Liver. *Arch Intern Med* **129**, 492–496 (1972).
19. Safra, N., Ling, G. V., Schaible, R. H. & Bannasch, D. L. Exclusion of Urate Oxidase as a Candidate Gene for Hyperuricosuria in the Dalmatian Dog Using an Interbreed Backcross. *J Hered* **96**, 750–754 (2005).
20. Bannasch, D. *et al.* Mutations in the SLC2A9 Gene Cause Hyperuricosuria and Hyperuricemia in the Dog. *PLoS Genet* **4**, (2008).
21. Oda, M., Satta, Y., Takenaka, O. & Takahata, N. Loss of Urate Oxidase Activity in Hominoids and its Evolutionary Implications. *Mol Biol Evol* **19**, 640–653 (2002).
22. Wu, X. W., Muzny, D. M., Lee, C. C. & Caskey, C. T. Two independent mutational events in the loss of urate oxidase during hominoid evolution. *J. Mol. Evol.* **34**, 78–84 (1992).
23. Ames, B. N., Cathcart, R., Schwiers, E. & Hochstein, P. Uric acid provides an antioxidant defense in humans against oxidant- and radical-caused aging and cancer: a hypothesis. *PNAS* **78**, 6858–6862 (1981).
24. Tan, P. K., Farrar, J. E., Gaucher, E. A. & Miner, J. N. Coevolution of URAT1 and Uricase during Primate Evolution: Implications for Serum Urate Homeostasis and Gout. *Mol Biol Evol* msw116 (2016). doi:10.1093/molbev/msw116
25. Kratzer, J. T. *et al.* Evolutionary history and metabolic insights of ancient mammalian uricases. *Proceedings of the National Academy of Sciences* **111**, 3763–3768 (2014).
26. De Vera, M. *et al.* Gout and the risk of Parkinson's disease: a cohort study. *Arthritis Rheum.* **59**, 1549–1554 (2008).
27. Alonso, A., Rodríguez, L. A. G., Logroscino, G. & Hernán, M. A. Gout and risk of Parkinson disease: a prospective study. *Neurology* **69**, 1696–1700 (2007).
28. Hooper, D. C. *et al.* Uric acid, a natural scavenger of peroxynitrite, in experimental allergic encephalomyelitis and multiple sclerosis. *Proc Natl Acad Sci U S A* **95**, 675–680 (1998).
29. Kuzkaya, N., Weissmann, N., Harrison, D. G. & Dikalov, S. Interactions of peroxynitrite with uric acid in the presence of ascorbate and thiols: implications for uncoupling endothelial nitric oxide synthase. *Biochem. Pharmacol.* **70**, 343–354 (2005).
30. Sautin, Y. Y. & Johnson, R. J. Uric acid: the oxidant-antioxidant paradox. *Nucleosides Nucleotides Nucleic Acids* **27**, 608–619 (2008).

31. Kanellis, J. *et al.* Uric acid stimulates monocyte chemoattractant protein-1 production in vascular smooth muscle cells via mitogen-activated protein kinase and cyclooxygenase-2. *Hypertension* **41**, 1287–1293 (2003).
32. Khosla, U. M. *et al.* Hyperuricemia induces endothelial dysfunction. *Kidney Int.* **67**, 1739–1742 (2005).
33. Gersch, C. *et al.* Inactivation of Nitric Oxide by Uric Acid. *Nucleosides Nucleotides Nucleic Acids* **27**, 967–978 (2008).
34. Sautin, Y. Y., Nakagawa, T., Zharikov, S. & Johnson, R. J. Adverse effects of the classic antioxidant uric acid in adipocytes: NADPH oxidase-mediated oxidative/nitrosative stress. *Am. J. Physiol., Cell Physiol.* **293**, C584-596 (2007).
35. Shi, Y., Mucsi, A. D. & Ng, G. Monosodium urate crystals in inflammation and immunity. *Immunol. Rev.* **233**, 203–217 (2010).
36. Eisenbacher, J. L. *et al.* S100A4 and uric acid promote mesenchymal stromal cell induction of IL-10+/IDO+ lymphocytes. *J. Immunol.* **192**, 6102–6110 (2014).
37. So, A. & Thorens, B. Uric acid transport and disease. *Journal of Clinical Investigation* **120**, 1791–1799 (2010).
38. Enomoto, A. *et al.* Molecular identification of a renal urate–anion exchanger that regulates blood urate levels. *Nature* **417**, 447–452 (2002).
39. Ekaratanawong, S. *et al.* Human organic anion transporter 4 is a renal apical organic anion/dicarboxylate exchanger in the proximal tubules. *J. Pharmacol. Sci.* **94**, 297–304 (2004).
40. Cha, S. H. *et al.* Molecular cloning and characterization of multispecific organic anion transporter 4 expressed in the placenta. *J. Biol. Chem.* **275**, 4507–4512 (2000).
41. Bahn, A. *et al.* Identification of a new urate and high affinity nicotinate transporter, hOAT10 (SLC22A13). *J. Biol. Chem.* **283**, 16332–16341 (2008).
42. Sekine, T., Watanabe, N., Hosoyamada, M., Kanai, Y. & Endou, H. Expression cloning and characterization of a novel multispecific organic anion transporter. *J. Biol. Chem.* **272**, 18526–18529 (1997).
43. Bakhiya, A., Bahn, A., Burckhardt, G. & Wolff, N. Human organic anion transporter 3 (hOAT3) can operate as an exchanger and mediate secretory urate flux. *Cell. Physiol. Biochem.* **13**, 249–256 (2003).
44. Kusuhara, H. *et al.* Molecular cloning and characterization of a new multispecific organic anion transporter from rat brain. *J. Biol. Chem.* **274**, 13675–13680 (1999).
45. Cha, S. H. *et al.* Identification and characterization of human organic anion transporter 3 expressing predominantly in the kidney. *Mol. Pharmacol.* **59**, 1277–1286 (2001).
46. Van Aubel, R. A. M. H., Smeets, P. H. E., van den Heuvel, J. J. M. W. & Russel, F. G. M. Human organic anion transporter MRP4 (ABCC4) is an efflux pump for the purine end metabolite urate with multiple allosteric substrate binding sites. *Am. J. Physiol. Renal Physiol.* **288**, F327-333 (2005).
47. Nakayama, A. *et al.* ABCG2 is a high-capacity urate transporter and its genetic impairment increases serum uric acid levels in humans. *Nucleosides Nucleotides Nucleic Acids* **30**, 1091–1097 (2011).
48. Huls, M. *et al.* The breast cancer resistance protein transporter ABCG2 is expressed in the human kidney proximal tubule apical membrane. *Kidney Int.* **73**, 220–225 (2008).
49. Xu, X., Li, C., Zhou, P. & Jiang, T. Uric acid transporters hiding in the intestine. *Pharmaceutical Biology* **54**, 3151–3155 (2016).
50. Bhatnagar, V. *et al.* Analysis of ABCG2 and other urate transporters in uric acid homeostasis in chronic kidney disease: potential role of remote sensing and signaling. *Clin Kidney J* **9**, 444–453 (2016).
51. Dehghan, A. *et al.* Association of three genetic loci with uric acid concentration and risk of gout: a genome-wide association study. *The Lancet* **372**, 1953–1961 (2008).
52. Jutabha, P. *et al.* Human sodium phosphate transporter 4 (hNPT4/SLC17A3) as a common renal secretory pathway for drugs and urate. *J. Biol. Chem.* **285**, 35123–35132 (2010).
53. Chiba, T. *et al.* NPT1/SLC17A1 is a renal urate exporter in humans and its common gain-of-function variant decreases the risk of renal underexcretion gout. *Arthritis & Rheumatology (Hoboken, N.J.)* **67**, 281–287 (2015).
54. Doege, H., Bocianski, A., Joost, H. G. & Schürmann, A. Activity and genomic organization of human glucose transporter 9 (GLUT9), a novel member of the family of sugar-transport facilitators predominantly expressed in brain and leucocytes. *Biochem. J.* **350 Pt 3**, 771–776 (2000).
55. Vitart, V. *et al.* SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nature Genetics* **40**, 437–442 (2008).
56. Döring, A. *et al.* SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. *Nature Genetics* **40**, 430 (2008).
57. Caulfield, M. J. *et al.* SLC2A9 Is a High-Capacity Urate Transporter in Humans. *PLOS Med* **5**, e197 (2008).
58. Augustin, R. *et al.* Identification and characterization of human glucose transporter-like protein-9 (GLUT9): alternative splicing alters trafficking. *J. Biol. Chem.* **279**, 16229–16236 (2004).
59. Bibert, S. *et al.* Mouse GLUT9: evidences for a urate uniporter. *Am. J. Physiol. Renal Physiol.* **297**, F612-619 (2009).
60. Matsuo, H. *et al.* Mutations in glucose transporter 9 gene SLC2A9 cause renal hypouricemia. *Am. J. Hum. Genet.* **83**, 744–751 (2008).
61. Lally, E. V., Ho, G. & Kaplan, S. R. The clinical spectrum of gouty arthritis in women. *Arch. Intern. Med.* **146**, 2221–2225 (1986).

62. Nicholls, A., Snaith, M. L. & Scott, J. T. Effect of oestrogen therapy on plasma and urinary levels of uric acid. *Br Med J* **1**, 449–451 (1973).
63. Sumino, H., Ichikawa, S., Kanda, T., Nakamura, T. & Sakamaki, T. Reduction of serum uric acid by hormone replacement therapy in postmenopausal women with hyperuricaemia. *Lancet* **354**, 650 (1999).
64. Yahyaoui, R. *et al.* Effect of Long-Term Administration of Cross-Sex Hormone Therapy on Serum and Urinary Uric Acid in Transsexual Persons. *J Clin Endocrinol Metab* **93**, 2230–2233 (2008).
65. Jung, J. H. *et al.* Serum uric acid levels and hormone therapy type: a retrospective cohort study of postmenopausal women. *Menopause* **25**, 77 (2018).
66. Li, S. *et al.* The GLUT9 Gene Is Associated with Serum Uric Acid Levels in Sardinia and Chianti Cohorts. *PLOS Genetics* **3**, e194 (2007).
67. Brandstätter, A. *et al*. Sex-specific association of the putative fructose transporter SLC2A9 variants with uric acid levels is modified by BMI. *Diabetes Care* **31**, 1662–1667 (2008).
68. Li, W.-D. *et al.* A genome wide association study of plasma uric acid levels in obese cases and never-overweight controls. *Obesity (Silver Spring)* **21**, E490-494 (2013).
69. Huffman, J. E. *et al.* Modulation of Genetic Associations with Serum Urate Levels by Body-Mass-Index in Humans. *PLOS ONE* **10**, e0119752 (2015).
70. Woodward, O. M. *et al.* Identification of a urate transporter, ABCG2, with a common functional polymorphism causing gout. *PNAS* **106**, 10338–10342 (2009).
71. Uchino, H. *et al.* p-aminohippuric acid transport at renal apical membrane mediated by human inorganic phosphate transporter NPT1. *Biochem. Biophys. Res. Commun.* **270**, 254–259 (2000).
72. Nakayama, A. *et al.* A common missense variant of monocarboxylate transporter 9 (MCT9/SLC16A9) gene is associated with renal overload gout, but not with all gout susceptibility. *Hum Cell* **26**, 133–136 (2013).
73. Taniguchi, A. & Kamatani, N. Control of renal uric acid excretion and gout. *Current Opinion in Rheumatology* **20**, 192–197 (2008).
74. Köttgen, A. *et al.* Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nature Genetics* **45**, 145–154 (2013).
75. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
76. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol* **27**, S2–S8 (2017).
77. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nature Genetics* **50**, 390–400 (2018).
78. Culleton, B. F., Larson, M. G., Kannel, W. B. & Levy, D. Serum uric acid and risk for cardiovascular disease and death: the Framingham Heart Study. *Ann. Intern. Med.* **131**, 7–13 (1999).
79. Mikkelsen, W. M., Dodge, H. J. & Valkenburg, H. The Distribution of Serum Uric Acid Values in a Population Unselected as to Gout or Hyperurieemia. *Journal of Occupational and Environmental Medicine* **8**, 48 (1966).
80. Zalokar, J., Lellouch, J., Claude, J. R. & Kuntz, D. Epidemiology of serum uric acid and gout in Frenchmen. *J Chronic Dis* **27**, 59–75 (1974).
81. Kuzuya, M., Ando, F., Iguchi, A. & Shimokata, H. Effect of Aging on Serum Uric Acid Levels: Longitudinal Changes in a Large Japanese Population Group. *J Gerontol A Biol Sci Med Sci* **57**, M660–M664 (2002).
82. Li, X. *et al.* Serum uric acid levels and multiple health outcomes: umbrella review of evidence from observational studies, randomised controlled trials, and Mendelian randomisation studies. *BMJ* **357**, j2376 (2017).
83. Borghi, C. Re: Uric acid: beyond the interpretation of serum levels. (2017).
84. Dalbeth, N., Merriman, T. R. & Stamp, L. K. Gout. *The Lancet* **388**, 2039–2052 (2016).
85. Martinon, F., Pétrilli, V., Mayor, A., Tardivel, A. & Tschopp, J. Gout-associated uric acid crystals activate the NALP3 inflammasome. *Nature* **440**, 237–241 (2006).
86. Cronstein, B. N. & Sunkureddi, P. Mechanistic aspects of inflammation and clinical management of inflammation in acute gouty arthritis. *J Clin Rheumatol* **19**, 19–29 (2013).
87. Hall, A. P., Barry, P. E., Dawber, T. R. & McNamara, P. M. Epidemiology of gout and hyperuricemia: A long-term population study. *The American Journal of Medicine* **42**, 27–37 (1967).
88. Kuo, C.-F., Grainge, M. J., Zhang, W. & Doherty, M. Global epidemiology of gout: prevalence, incidence and risk factors. *Nat Rev Rheumatol* **11**, 649–662 (2015).
89. Wallace, K. L., Riedel, A. A., Joseph-Ridge, N. & Wortmann, R. Increasing prevalence of gout and hyperuricemia over 10 years among older adults in a managed care population. *J. Rheumatol.* **31**, 1582–1587 (2004).
90. Zhu, Y., Pandya, B. J. & Choi, H. K. Prevalence of gout and hyperuricemia in the US general population: the National Health and Nutrition Examination Survey 2007-2008. *Arthritis Rheum.* **63**, 3136–3141 (2011).
91. Kuo, C.-F., Grainge, M. J., Mallen, C., Zhang, W. & Doherty, M. Rising burden of gout in the UK but continuing suboptimal management: a nationwide population study. *Annals of the Rheumatic Diseases* **74**, 661–667 (2015).
92. Takahashi, S. *et al.* Close correlation between visceral fat accumulation and uric acid metabolism in healthy men. *Metab. Clin. Exp.* **46**, 1162–1165 (1997).
93. Kwon, H. & Pessin, J. E. Adipokines Mediate Inflammation and Insulin Resistance. *Front Endocrinol (Lausanne)* **4**, (2013).
94. Fuentes, E., Fuentes, F., Vilahur, G., Badimon, L. & Palomo, I. Mechanisms of chronic state of inflammation as mediators that link obese adipose tissue and metabolic syndrome. *Mediators Inflamm.* **2013**, 136584 (2013).

95. Battelli, M. G., Bortolotti, M., Polito, L. & Bolognesi, A. The role of xanthine oxidoreductase and uric acid in metabolic syndrome. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1864**, 2557–2565 (2018).
96. Toyoki, D. *et al.* Insulin stimulates uric acid reabsorption via regulating urate transporter 1 and ATP-binding cassette subfamily G member 2. *Am. J. Physiol. Renal Physiol.* **313**, F826–F834 (2017).
97. Cai, W. *et al.* Uric Acid Induces Endothelial Dysfunction by Activating the HMGB1/RAGE Signaling Pathway. *Biomed Res Int* **2017**, 4391920 (2017).
98. Choi, Y.-J. *et al.* Uric acid induces endothelial dysfunction by vascular insulin resistance associated with the impairment of nitric oxide synthesis. *The FASEB Journal* **28**, 3197–3204 (2014).
99. Borgi, L. *et al.* Effect of Uric Acid-Lowering Agents on Endothelial Function: A Randomized, Double-Blind, Placebo-Controlled Trial. *Hypertension* **69**, 243–248 (2017).
100. Kutzing, M. K. & Firestein, B. L. Altered Uric Acid Levels and Disease States. *Journal of Pharmacology and Experimental Therapeutics* **324**, 1–7 (2007).
101. Feig, D. I., Soletsky, B. & Johnson, R. J. Effect of allopurinol on blood pressure of adolescents with newly diagnosed essential hypertension: a randomized trial. *JAMA* **300**, 924–932 (2008).
102. Feig, D. I. & Johnson, R. J. Hyperuricemia in Childhood Primary Hypertension. *Hypertension* **42**, 247–252 (2003).
103. Borghi, C. *et al.* Serum uric acid levels are associated with cardiovascular risk score: A post hoc analysis of the EURIKA study. *International Journal of Cardiology* **253**, 167–173 (2018).
104. Carluccio, E., Coiro, S. & Ambrosio, G. Unraveling the relationship between serum uric acid levels and cardiovascular risk. *International Journal of Cardiology* **253**, 174–175 (2018).
105. Berger, L. & Yü, T. F. Renal function in gout. IV. An analysis of 524 gouty subjects including long-term follow-up studies. *Am. J. Med.* **59**, 605–613 (1975).
106. Saito, I. *et al.* Serum uric acid and the renin-angiotensin system in hypertension. *J Am Geriatr Soc* **26**, 241–247 (1978).
107. Sánchez-Lozada, L. G. *et al.* Treatment with the xanthine oxidase inhibitor febuxostat lowers uric acid and alleviates systemic and glomerular hypertension in experimental hyperuricaemia. *Nephrol. Dial. Transplant.* **23**, 1179–1185 (2008).
108. Li, L. *et al.* Is hyperuricemia an independent risk factor for new-onset chronic kidney disease?: A systematic review and meta-analysis based on observational cohort studies. *BMC Nephrol* **15**, 122 (2014).
109. Bose, B. *et al.* Effects of uric acid-lowering therapy on renal outcomes: a systematic review and meta-analysis. *Nephrol Dial Transplant* **29**, 406–413 (2014).
110. Kang, D.-H. *et al.* A role for uric acid in the progression of renal disease. *J. Am. Soc. Nephrol.* **13**, 2888–2897 (2002).
111. Auinger, P., Kieburtz, K. & Mcdermott, M. P. The relationship between uric acid levels and Huntington's disease progression. *Mov. Disord.* **25**, 224–228 (2010).
112. Lu, N. *et al.* Gout and the risk of Alzheimer's disease: a population-based, BMI-matched cohort study. *Ann Rheum Dis* annrheumdis-2014-206917 (2015). doi:10.1136/annrheumdis-2014-206917
113. Knapp, C. M. *et al.* Serum uric acid levels in optic neuritis. *Mult. Scler.* **10**, 278–280 (2004).
114. Munan, L., Kelly, A. & Petitclerc, C. Population serum urate levels and their correlates. The Sherbrooke regional study. *Am. J. Epidemiol.* **103**, 369–382 (1976).
115. Folkersen, L. *et al.* Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLOS Genetics* **13**, e1006706 (2017).
116. Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature Communications* **9**, 3268 (2018).
117. Carayol, J. *et al.* Protein quantitative trait locus study in obesity during weight-loss identifies a leptin regulator. *Nature Communications* **8**, 2084 (2017).
118. Demirkan, A. *et al.* Genome-Wide Association Study Identifies Novel Loci Associated with Circulating Phospho- and Sphingolipid Concentrations. *PLOS Genetics* **8**, e1002490 (2012).
119. Rhee, E. P. *et al.* A Genome-wide Association Study of the Human Metabolome in a Community-Based Cohort. *Cell Metabolism* **18**, 130–143 (2013).
120. Albrecht, E. *et al.* Metabolite profiling reveals new insights into the regulation of serum urate in humans. *Metabolomics* **10**, (2014).
121. Vitart, V. *et al.* 3000 years of solitude: extreme differentiation in the island isolates of Dalmatia, Croatia. *European Journal of Human Genetics* **14**, 478–487 (2006).
122. Rudan, I. *et al.* "10 001 Dalmatians:" Croatia Launches Its National Biobank. *Croatian Medical Journal* **50**, 4–6 (2009).
123. Di Angelantonio, E. *et al.* Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. *The Lancet* **390**, 2360–2371 (2017).
124. Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol* **44**, 1137–1147 (2015).
125. Tigchelaar, E. F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).

126. Krakauer, N. Y. & Krakauer, J. C. A New Body Shape Index Predicts Mortality Hazard Independently of Body Mass Index. *PLOS ONE* **7**, e39504 (2012).
127. Assarsson, E. *et al.* Homogenous 96-Plex PEA Immunoassay Exhibiting High Sensitivity, Specificity, and Excellent Scalability. *PLOS ONE* **9**, e95192 (2014).
128. Liebisch, G. *et al.* Quantitative measurement of different ceramide species from crude cellular extracts by electrospray ionization tandem mass spectrometry (ESI-MS/MS). *J. Lipid Res.* **40**, 1539–1546 (1999).
129. Liebisch, G., Lieser, B., Rathenberg, J., Drobnik, W. & Schmitz, G. High-throughput quantification of phosphatidylcholine and sphingomyelin by electrospray ionization tandem mass spectrometry coupled with isotope correction algorithm. *Biochim. Biophys. Acta* **1686**, 108–117 (2004).
130. Levey, A. S. *et al.* Using Standardized Serum Creatinine Values in the Modification of Diet in Renal Disease Study Equation for Estimating Glomerular Filtration Rate. *Annals of Internal Medicine* **145**, 247 (2006).
131. Levey, A. S. *et al*. A new equation to estimate glomerular filtration rate. *Ann. Intern. Med.* **150**, 604–612 (2009).
132. Aulchenko, Y. S., Ripke, S., Isaacs, A. & Duijn, C. M. van. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
133. Harrel, F. E. *Hmisc: Harrell Miscellaneous. R package version 3.17-4. https://CRAN.R-project.org/package=Hmisc*. (2016).
134. Kim, S. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Commun Stat Appl Methods* **22**, 665–674 (2015).
135. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**, 1–9 (2006).
136. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
137. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
138. Eden, E., Lipson, D., Shannon, P. & Yakhini, Z. Discovering Motifs in Ranked Lists of DNA Sequences. *PLOS Computational Biology* **3**, e39 (2007).
139. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1–22 (2010).
140. Staley, J. R. *et al.* PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).
141. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**, D896–D901 (2017).
142. Leslie, R., O'Donnell, C. J. & Johnson, A. D. GRASP: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* **30**, i185–i194 (2014).
143. Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics* **39**, 1181–1186 (2007).
144. Johnson, T. *Genetics ToolboX*. (2015).
145. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics* **10**, e1004383 (2014).
146. McKeigue, P., Colombo, M. & Spiliopoulou, A. GENOSCORES: a platform for calculating genotypic predictors of binary and quantitative phenotypes. Available at: https://pm2.phs.ed.ac.uk/genoscores/. (Accessed: 11th January 2019)
147. Consortium, T. 1000 G. P. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
148. Bulik-Sullivan, B. *et al.* An Atlas of Genetic Correlations across Human Diseases and Traits. *Nat Genet* **47**, 1236–1241 (2015).
149. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).
150. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
151. Melzer, D. *et al.* A Genome-Wide Association Study Identifies Protein Quantitative Trait Loci (pQTLs). *PLOS Genetics* **4**, e1000072 (2008).
152. Garge, N. *et al.* Identification of Quantitative Trait Loci Underlying Proteome Variation in Human Lymphoblastoid Cells. *Mol Cell Proteomics* **9**, 1383–1399 (2010).
153. Lourdusamy, A. *et al.* Identification of cis-regulatory variation influencing protein abundance levels in human plasma. *Hum Mol Genet* **21**, 3719–3726 (2012).
154. Choi, S. H. *et al.* Six Novel Loci Associated with Circulating VEGF Levels Identified by a Meta-analysis of Genome-Wide Association Studies. *PLOS Genetics* **12**, e1005874 (2016).
155. Beenken, A. & Mohammadi, M. The FGF family: biology, pathophysiology and therapy. *Nature Reviews Drug Discovery* **8**, 235–253 (2009).
156. He, X. *et al.* The association of serum FGF23 and non-alcoholic fatty liver disease is independent of vitamin D in type 2 diabetes patients. *Clin. Exp. Pharmacol. Physiol.* **45**, 668–674 (2018).

157. Shih, M.-H. *et al.* Association between serum uric acid and nonalcoholic fatty liver disease in the US population. *Journal of the Formosan Medical Association* **114**, 314–320 (2015).
158. Li, Y., Xu, C., Yu, C., Xu, L. & Miao, M. Association of serum uric acid level with non-alcoholic fatty liver disease: A cross-sectional study. *Journal of Hepatology* **50**, 1029–1034 (2009).
159. Sugiura, H. *et al.* Fibroblast growth factor 23 is upregulated in the kidney in a chronic kidney disease rat model. *PLOS ONE* **13**, e0191706 (2018).
160. Fayed, A. *et al.* Fibroblast growth factor-23 is a strong predictor of insulin resistance among chronic kidney disease patients. *Renal Failure* **40**, 226–230 (2018).
161. Gutiérrez, O. M., Wolf, M. & Taylor, E. N. Fibroblast Growth Factor 23, Cardiovascular Disease Risk Factors, and Phosphorus Intake in the Health Professionals Follow-up Study. *CJASN* **6**, 2871–2878 (2011).
162. Sakoh, T. *et al.* Associations of fibroblast growth factor 23 with urate metabolism in patients with chronic kidney disease. *Metabolism* **65**, 1498–1507 (2016).
163. Bacchetta, J., Cochat, P., Salusky, I. B. & Wesseling-Perry, K. Uric acid and IGF1 as possible determinants of FGF23 metabolism in children with normal renal function. *Pediatr Nephrol* **27**, 1131–1138 (2012).
164. Kharitonenkov, A. *et al.* FGF-21 as a novel metabolic regulator. *J Clin Invest* **115**, 1627–1635 (2005).
165. Zhang, J. *et al.* The role of FGF21 in type 1 diabetes and its complications. *Int J Biol Sci* **14**, 1000–1011 (2018).
166. Cuevas-Ramos, D. *et al.* Daily physical activity, fasting glucose, uric acid, and body mass index are independent factors associated with serum fibroblast growth factor 21 levels. *European Journal of Endocrinology* **163**, 469–477 (2010).
167. Cheung, C. Y. Y. *et al.* An Exome-Chip Association Analysis in Chinese Subjects Reveals a Functional Missense Variant of GCKR That Regulates FGF21 Levels. *Diabetes* **66**, 1723–1728 (2017).
168. Rees, M. G. *et al.* Cellular characterisation of the GCKR P446L variant associated with type 2 diabetes risk. *Diabetologia* **55**, 114–122 (2012).
169. Saxena, R. *et al.* Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels. *Science* **316**, 1331–1336 (2007).
170. Orho-Melander, M. *et al.* Common Missense Variant in the Glucokinase Regulatory Protein Gene Is Associated With Increased Plasma Triglyceride and C-Reactive Protein but Lower Fasting Glucose Concentrations. *Diabetes* **57**, 3112–3121 (2008).
171. Kolz, M. *et al.* Meta-Analysis of 28,141 Individuals Identifies Common Variants within Five New Loci That Influence Uric Acid Concentrations. *PLOS Genetics* **5**, e1000504 (2009).
172. Ding, H., Kharboutli, M., Saxena, R. & Wu, T. Insulin-like growth factor binding protein-2 as a novel biomarker for disease activity and renal pathology changes in lupus nephritis. *Clinical & Experimental Immunology* **184**, 11–18 (2016).
173. Frystyk, J., Skjaerbaek, C., Vestbo, E., Fisker, S. & Orskov, H. Circulating levels of free insulin-like growth factors in obese subjects: the impact of type 2 diabetes. *Diabetes Metab. Res. Rev.* **15**, 314–322 (1999).
174. Nam, S. Y. *et al.* Effect of obesity on total and free insulin-like growth factor (IGF)-1, and their relationship to IGF-binding protein (BP)-1, IGFBP-2, IGFBP-3, insulin, and growth hormone. *Int. J. Obes. Relat. Metab. Disord.* **21**, 355–359 (1997).
175. Heald, A. H. *et al.* Insulin-like growth factor binding protein-2 (IGFBP-2) is a marker for the metabolic syndrome. *Exp. Clin. Endocrinol. Diabetes* **114**, 371–376 (2006).
176. Xi, G., Wai, C., DeMambro, V., Rosen, C. J. & Clemmons, D. R. IGFBP-2 Directly Stimulates Osteoblast Differentiation. *Journal of Bone and Mineral Research* **29**, 2427–2438 (2014).
177. The Human Protein Atlas. Available at: https://www.proteinatlas.org/. (Accessed: 2nd November 2018)
178. Enroth, S., Johansson, Å., Enroth, S. B. & Gyllensten, U. Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. *Nature Communications* **5**, 4684 (2014).
179. Ferrer-Admetlla, A. *et al.* A Natural History of FUT2 Polymorphism in Humans. *Mol Biol Evol* **26**, 1993–2003 (2009).
180. Prins, B. P. *et al.* Genome-wide analysis of health-related biomarkers in the UK Household Longitudinal Study reveals novel associations. *Scientific Reports* **7**, 11008 (2017).
181. Brand, F. N., McGee, D. L., Kannel, W. B., Stokes, J. & Castelli, W. P. Hyperuricemia as a risk factor of coronary heart disease: The Framingham Study. *Am. J. Epidemiol.* **121**, 11–18 (1985).
182. Zhang, J., Zhang, Y., Wu, Q. & Chen, B. Uric acid induces oxidative stress via an activation of the renin–angiotensin system in 3T3-L1 adipocytes. *Endocrine* **48**, 135–142 (2015).
183. Walker, W. G., Whelton, P. K., Saito, H., Russell, R. P. & Hermann, J. Relation between blood pressure and renin, renin substrate, angiotensin II, aldosterone and urinary sodium and potassium in 574 ambulatory subjects. *Hypertension* **1**, 287–291 (1979).
184. Ferris, T. F. & Gorden, P. Effect of angiotensin and norepinephrine upon urate clearance in man. *The American Journal of Medicine* **44**, 359–365 (1968).
185. Lee, J. J. *et al.* Relationship between uric acid and blood pressure in different age groups. *Clinical Hypertension* **21**, (2015).
186. Zheng, H., Li, N., Ding, Y. & Miao, P. Losartan alleviates hyperuricemia-induced atherosclerosis in a rabbit model. *Int J Clin Exp Pathol* **8**, 10428–10435 (2015).

187. Hamada, T. *et al.* Uricosuric Action of Losartan via the Inhibition of Urate Transporter 1 (URAT 1) in Hypertensive Patients. *Am J Hypertens* **21**, 1157–1162 (2008).
188. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707 (2010).
189. Shum, B. O. V. *et al.* The adipocyte fatty acid–binding protein aP2 is required in allergic airway inflammation. *J Clin Invest* **116**, 2183–2192 (2006).
190. Tsushima, Y. *et al.* Uric Acid Secretion from Adipose Tissue and Its Increase in Obesity. *J Biol Chem* **288**, 27138–27149 (2013).
191. Yan, J.-Y. & Wang, X.-J. Expression and significance of adipocyte fatty acid-binding protein in placenta, serum and umbilical cord blood in preeclampsia. *Zhonghua Fu Chan Ke Za Zhi* **45**, 885–890 (2010).
192. Cai, H. *et al.* Benzbromarone, an old uricosuric drug, inhibits human fatty acid binding protein 4 in vitro and lowers the blood glucose level in db/db mice. *Acta Pharmacol Sin* **34**, 1397–1402 (2013).
193. Inokuchi, T. *et al.* Effects of Benzbromarone and Allopurinol on Adiponectin In Vivo and In Vitro. *Horm Metab Res* **41**, 327–332 (2009).
194. Crișan, T. O. *et al.* Soluble uric acid primes TLR-induced proinflammatory cytokine production by human primary cells via inhibition of IL-1Ra. *Annals of the Rheumatic Diseases* **75**, 755–762 (2016).
195. Kobayashi, T., Kouzaki, H. & Kita, H. Human Eosinophils Recognize Endogenous Danger Signal Crystalline Uric Acid and Produce Proinflammatory Cytokines Mediated by Autocrine ATP. *The Journal of Immunology* **184**, 6350–6358 (2010).
196. Borges, F. T., Dalboni, M. A., Michelacci, Y. M. & Schor, N. Noncrystalline uric acid inhibits proteoglycan and glycosaminoglycan synthesis in distal tubular epithelial cells (MDCK). *Brazilian Journal of Medical and Biological Research* **43**, 957–963 (2010).
197. Rull, A. *et al.* Serum paraoxonase-3 concentration is associated with insulin sensitivity in peripheral artery disease and with inflammation in coronary artery disease. *Atherosclerosis* **220**, 545–551 (2012).
198. Kirschbaum, B. Correlation studies of plasma paraoxonase activity and uric acid concentration with AAPH-Induced erythrocyte hemolysis in hemodialysis patients. *Artif Organs* **28**, 259–64 (2004).
199. Mapstone, M. *et al.* Plasma phospholipids identify antecedent memory impairment in older adults. *Nature Medicine* **20**, 415–418 (2014).
200. Mastrokolias, A. *et al.* Integration of targeted metabolomics and transcriptomics identifies deregulation of phosphatidylcholine metabolism in Huntington's disease peripheral blood samples. *Metabolomics* **12**, (2016).
201. Figarska Sylwia M. *et al.* Associations of Circulating Protein Levels With Lipid Fractions in the General Population. *Arteriosclerosis, Thrombosis, and Vascular Biology* **38**, 2505–2518 (2018).
202. Rasheed, H., Hughes, K., Flynn, T. J. & Merriman, T. R. Mendelian Randomization Provides No Evidence for a Causal Role of Serum Urate in Increasing Serum Triglyceride Levels. *Circ Cardiovasc Genet* **7**, 830–837 (2014).
203. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
204. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *bioRxiv* 035170 (2015). doi:10.1101/035170
205. Vries, P. S. de *et al.* Comparison of HapMap and 1000 Genomes Reference Panels in a Large-Scale Genome-Wide Association Study. *PLOS ONE* **12**, e0167742 (2017).
206. Nagy, R. *et al.* Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants. *Genome Medicine* **9**, 23 (2017).
207. Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., Duijn, C. M. van & Aulchenko, Y. S. Rapid variance components–based method for whole-genome association analysis. *Nature Genetics* **44**, 1166–1170 (2012).
208. Haller, T., Kals, M., Esko, T., Mägi, R. & Fischer, K. RegScan: a GWAS tool for quick estimation of allele effects on continuous traits and their combinations. *Brief Bioinform* bbt066 (2013). doi:10.1093/bib/bbt066
209. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
210. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122 (2016).
211. Church, W. H. & Ward, V. L. Uric acid is reduced in the substantia nigra in parkinson's disease: Effect on dopamine oxidation. *Brain Research Bulletin* **33**, 419–425 (1994).
212. Cipriani, S., Chen, X. & Schwarzschild, M. A. Urate: a novel biomarker of Parkinson's disease risk, diagnosis and prognosis. *Biomark Med* **4**, 701–712 (2010).
213. Lessard, C. J. *et al.* Identification of a Systemic Lupus Erythematosus Risk Locus Spanning ATG16L2, FCHSD2, and P2RY2 in Koreans. *Arthritis & Rheumatology (Hoboken, N.J.)* **68**, 1197–1209 (2016).
214. Calich, A. L. *et al.* Serum uric acid levels are associated with lupus nephritis in patients with normal renal function. *Clin. Rheumatol.* **37**, 1223–1228 (2018).
215. Fuyuno, Y. *et al.* Genetic characteristics of inflammatory bowel disease in a Japanese population. *J. Gastroenterol.* **51**, 672–681 (2016).
216. Fan, X. *et al.* Identification of the gene encoding the enzyme deficient in mucopolysaccharidosis IIIC (Sanfilippo disease type C). *Am. J. Hum. Genet.* **79**, 738–744 (2006).
217. Hrebícek, M. *et al.* Mutations in TMEM76* cause mucopolysaccharidosis IIIC (Sanfilippo C syndrome). *Am. J. Hum. Genet.* **79**, 807–819 (2006).

218. Haer-Wigman, L. *et al.* Non-syndromic retinitis pigmentosa due to mutations in the mucopolysaccharidosis type IIIC gene, heparan-alpha-glucosaminide N-acetyltransferase (HGSNAT). *Hum. Mol. Genet.* **24**, 3742–3751 (2015).
219. Moncrieff, C. L., Bailey, M. E. S., Morrison, N. & Johnson, K. J. Cloning and Chromosomal Localization of Human Cdc42-Binding Protein Kinase β. *Genomics* **57**, 297–300 (1999).
220. Kim, Y. J. *et al.* Overexpression and unique rearrangement of VH2 transcripts in immunoglobulin variable heavy chain genes in ankylosing spondylitis patients. *Experimental & Molecular Medicine* **42**, 319–326 (2010).
221. Kang, K. Y., Hong, Y. S., Park, S.-H. & Ju, J. H. Low Levels of Serum Uric Acid Increase the Risk of Low Bone Mineral Density in Young Male Patients with Ankylosing Spondylitis. *The Journal of Rheumatology* **42**, 968–974 (2015).
222. Mariani, E. *et al.* Meta-Analysis of Parkinson's Disease Transcriptome Data Using TRAM Software: Whole Substantia Nigra Tissue and Single Dopamine Neuron Differential Gene Expression. *PLoS One* **11**, (2016).
223. Lee, H.-K. *et al.* Dynamic Ca2+-Dependent Stimulation of Vesicle Fusion by Membrane-Anchored Synaptotagmin 1. *Science* **328**, 760–763 (2010).
224. Pattaro, C. *et al.* A meta-analysis of genome-wide data from five European isolates reveals an association of COL22A1, SYT1, and GABRR2 with serum creatinine level. *BMC Med. Genet.* **11**, 41 (2010).
225. Fuchsberger, C., Taliun, D., Pramstaller, P. P. & Pattaro, C. GWAtoolbox: an R package for fast quality control and handling of genome-wide association studies meta-analysis data. *Bioinformatics* **28**, 444–445 (2012).
226. Devlin, B., Roeder, K. & Wasserman, L. Genomic Control, a New Approach to Genetic-Based Association Studies. *Theoretical Population Biology* **60**, 155–166 (2001).
227. Dadd, T., Weale, M. E. & Lewis, C. M. A critical evaluation of genomic control methods for genetic association studies. *Genetic Epidemiology* **33**, 290–298 (2009).
228. Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nature Genetics* **36**, 512–517 (2004).
229. Wang, S. *et al.* Double genomic control is not effective to correct for population stratification in meta-analysis for genome-wide association studies. *Front. Genet.* **3**, (2012).
230. Mägi, R. *et al.* Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum Mol Genet* **26**, 3639–3650 (2017).
231. Morris, A. P. Transethnic meta-analysis of genomewide association studies. *Genetic Epidemiology* **35**, 809–822 (2011).
232. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-S3 (2012).
233. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am J Hum Genet* **88**, 76–82 (2011).
234. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* **12**, e1001779 (2015).
235. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications* **6**, 5890 (2015).
236. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, (2015).
237. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**, D109–D114 (2012).
238. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).
239. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* **39**, D691–D697 (2011).
240. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology* **25**, 309–316 (2007).
241. Blake, J. A., Bult, C. J., Eppig, J. T., Kadin, J. A. & Richardson, J. E. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res* **42**, D810–D817 (2014).
242. Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nature Genetics* **47**, 115–125 (2015).
243. Frey, B. J. & Dueck, D. Clustering by Passing Messages Between Data Points. *Science* **315**, 972–976 (2007).
244. Bodenhofer, U., Kothmeier, A. & Hochreiter, S. APCluster: an R package for affinity propagation clustering. *Bioinformatics* **27**, 2463–2464 (2011).
245. Watanabe, K., Taskesen, E., Bochoven, A. van & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications* **8**, 1826 (2017).
246. Kamatani, Y. *et al.* Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat. Genet.* **42**, 210–215 (2010).
247. Yang, Q. *et al.* Multiple Genetic Loci Influence Serum Urate Levels and Their Relationship With Gout and Cardiovascular Disease Risk Factors. *Circulation: Cardiovascular Genetics* **3**, 523–530 (2010).
248. Tin, A. *et al.* Genome-wide association study for serum urate concentrations and gout among African Americans identifies genomic risk loci and a novel URAT1 loss-of-function allele. *Hum Mol Genet* **20**, 4056–4068 (2011).

249. Okada, Y. *et al.* Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations. *Nat. Genet.* **44**, 904–909 (2012).
250. Merriman, T. R. Population Heterogeneity in the Genetic Control of Serum Urate. *Seminars in Nephrology* **31**, 420–425 (2011).
251. Ketharnathan, S. *et al.* A non-coding genetic variant maximally associated with serum urate levels is functionally linked to HNF4A-dependent PDZK1 expression. *Hum Mol Genet* **27**, 3964–3973 (2018).
252. Simmonds, H. A. *et al.* Polynesian women are also at risk for hyperuricaemia and gout because of a genetic defect in renal urate handling. *Rheumatology (Oxford)* **33**, 932–937 (1994).
253. Gibson, T., Waterworth, R., Hatfield, P., Robinson, G. & Bremner, K. Hyperuricaemia, gout and kidney function in New Zealand Maori men. *Rheumatology (Oxford)* **23**, 276–282 (1984).
254. Phipps-Green, A. J. *et al.* A strong role for the ABCG2 gene in susceptibility to gout in New Zealand Pacific Island and Caucasian, but not Māori, case and control sample sets. *Hum Mol Genet* **19**, 4813–4819 (2010).
255. Thorens, B. GLUT2, glucose sensing and glucose homeostasis. *Diabetologia* **58**, 221–232 (2015).
256. Sharma, M. *et al.* hZimp10 is an androgen receptor co-activator and forms a complex with SUMO-1 at replication foci. *EMBO J* **22**, 6101–6114 (2003).
257. Fewings, N. L. *et al.* The autoimmune risk gene ZMIZ1 is a vitamin D responsive marker of a molecular phenotype of multiple sclerosis. *Journal of Autoimmunity* **78**, 57–69 (2017).
258. Massa, J., O'Reilly, E., Munger, K. L., DeLorenze, G. N. & Ascherio, A. Serum uric acid and risk of multiple sclerosis. *J Neurol* **256**, 1643–1648 (2009).
259. Tay, C. G. *et al.* Succinic Semialdehyde Dehydrogenase Deficiency in a Chinese Boy: A Novel ALDH5A1 Mutation With Severe Phenotype. *J Child Neurol* **30**, 927–931 (2015).
260. Yu, D. *et al.* Modulation of ALDH5A1 and SLC22A7 by microRNA hsa-miR-29a-3p in human liver cells. *Biochemical Pharmacology* **98**, 671–680 (2015).
261. Yang, A. W., Sachs, A. J. & Nystuen, A. M. Deletion of Inpp5a causes ataxia and cerebellar degeneration in mice. *Neurogenetics* **16**, 277–285 (2015).
262. Xing, J. *et al.* TRIM29 promotes DNA virus infections by inhibiting innate immune response. *Nature Communications* **8**, 945 (2017).
263. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142-147 (2014).
264. Kumar, S., Ambrosini, G. & Bucher, P. SNP2TFBS – a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res* **45**, D139–D144 (2017).
265. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930–D934 (2012).
266. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* **17**, 405–423 (2015).
267. Widén, E. & Ripatti, S. Assessment of multifactorial coronary artery disease by utilizing genomic data. *Duodecim* **133**, 776–781 (2017).
268. Cutler, R. G. Urate and ascorbate: their possible roles as antioxidants in determining longevity of mammalian species. *Archives of Gerontology and Geriatrics* **3**, 321–348 (1984).
269. Lai, J. Y.-C. *et al.* Family History of Exceptional Longevity is Associated with Lower Serum Uric Acid Levels in Ashkenazi Jews. *J Am Geriatr Soc* **60**, 745–750 (2012).
270. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
271. Peedicayil, J. & Grayson, D. R. An epigenetic basis for an omnigenic model of psychiatric disorders. *Journal of Theoretical Biology* **443**, 52–55 (2018).
272. van der Harst Pim & Verweij Niek. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circulation Research* **122**, 433–443 (2018).
273. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. *Cell* **173**, 1573–1580 (2018).
274. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nature Genetics* **50**, 1593 (2018).
275. Global Biobank Engine, Stanford, CA. (2018). Available at: http://gbe.stanford.edu. (Accessed: 30th November 2018)
276. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* **70**, 214–223 (2016).
277. Zhu, Y., Pandya, B. J. & Choi, H. K. Comorbidities of Gout and Hyperuricemia in the US General Population: NHANES 2007-2008. *The American Journal of Medicine* **125**, 679-687.e1 (2012).
278. Warren, H. R. *et al.* Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nature Genetics* **49**, 403–415 (2017).
279. McLeod, A. I. & Changjiang, X. *bestglm.* (2018).
280. Schwartz, S. & Susser, E. Genome-Wide Association Studies: Does Only Size Matter? *American Journal of Psychiatry* (2010). doi:10.1176/appi.ajp.2010.10030465

281. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).
282. Kerr, S. M. *et al.* Electronic health record and genome-wide genetic data in Generation Scotland participants. *Wellcome Open Res* **2**, 85 (2017).
283. Howard, D. M. *et al.* Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. *Nature Communications* **9**, 1470 (2018).
284. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* **186**, 1026–1034 (2017).
285. Bustamante, C. D., De La Vega, F. M. & Burchard, E. G. Genomics for the world. *Nature* **475**, 163–165 (2011).
286. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature News* **538**, 161 (2016).
287. Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics* **25**, 489–494 (2009).
288. Carlson, C. S. *et al.* Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. *PLOS Biology* **11**, e1001661 (2013).
289. O'Connor, L. J. & Price, A. L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nature Genetics* **50**, 1728 (2018).
290. Siemons, L. & Krishnan, E. A Prediction Tool for Incident Gout Among Those With Hyperuricemia. in *ACR Meeting Abstracts* (2013).

XI. Supplementary Tables

Supplementary Table 1 – List of all Olink proteins and non-Olink phenotypes included in correlation and lasso regression analyses.

Figure Display Name	Protein Name	Symbol	UniProt ID
CVD2_101_BMP.6	Bone morphogenetic protein 6	BMP-6	P22004
CVD2_102_ANG.1	Angiopoietin-1	ANG-1	Q15389
CVD2_103_ADM	ADM	ADM	P35318
CVD2_105_CD40.L	CD40 ligand	CD40-L	P29965
CVD2_106_SLAMF7	SLAM family member 7	SLAMF7	Q9NQ25
CVD2_107_PlGF	Placenta growth factor	PGF	P49763
CVD2_108_ADAM.TS13	A disintegrin and metalloproteinase with thrombospondin motifs 13	ADAM-TS13	Q76LX8
CVD2_109_Protein.BOC	Brother of CDO	Protein BOC	Q9BWV1
CVD2_110_IL.4RA	Interleukin-4 receptor subunit alpha	IL-4RA	P24394
CVD2_111_SRC	Proto-oncogene tyrosine-protein kinase Src	SRC	P12931
CVD2_112_IL.1ra	Interleukin-1 receptor antagonist protein	IL-1ra	P18510
CVD2_113_IL.6	Interleukin-6	IL6	P05231
CVD2_114_TNFRSF10A	Tumour necrosis factor receptor superfamily member 10A	TNFRSF10A	O00220
CVD2_115_STK4	Serine/threonine-protein kinase 4	STK4	Q13043
CVD2_116_IDUA	Alpha-L-iduronidase	IDUA	P35475
CVD2_117_TNFRSF11A	Tumour necrosis factor receptor superfamily member 11A	TNFRSF11A	Q9Y6Q6
CVD2_118_PAR.1	Proteinase-activated receptor 1	PAR-1	P25116
CVD2_120_TRAIL.R2	TNF-related apoptosis-inducing ligand receptor 2	TRAIL-R2	O14763
CVD2_121_PRSS27	Serine protease 27	PRSS27	Q9BQR3
CVD2_122_TIE2	Angiopoietin-1 receptor	TIE2	Q02763
CVD2_123_TF	Tissue factor	TF	P13726
CVD2_124_IL1RL2	Interleukin-1 receptor-like 2	IL1RL2	Q9HB29
CVD2_125_PDGF.subunit.B	Platelet-derived growth factor subunit B	PDGF subunit B	P01127
CVD2_126_IL27	Interleukin-27	IL-27	Q8NEV9
CVD2_127_IL.17D	Interleukin-17D	IL-17D	Q8TAD2
CVD2_128_CXCL1	C-X-C motif chemokine 1	CXCL1	P09341
CVD2_129_LOX.1	Lectin-like oxidized LDL receptor 1	LOX-1	P78380
CVD2_130_Gal.9	Galectin-9	Gal-9	O00182
CVD2_131_GIF	Gastric intrinsic factor	GIF	P27352
CVD2_132_SCF	Stem cell factor	SCF	P21583
CVD2_133_IL.18	Interleukin-18	IL-18	Q14116
CVD2_134_FGF.21	Fibroblast growth factor 21	FGF-21	Q9NSA1
CVD2_135_PIgR	Polymeric immunoglobulin receptor	PIgR	P01833
CVD2_136_RAGE	Receptor for advanced glycosylation end products	RAGE	Q15109
CVD2_137_SOD2	Superoxide dismutase [Mn], mitochondrial	SOD2	P04179
CVD2_138_CTRC	Chymotrypsin C	CTRC	Q99895
CVD2_139_FGF.23	Fibroblast growth factor 23	FGF-23	Q9GZV9
CVD2_140_SPON2	Spondin-2	SPON2	Q9BUD6
CVD2_141_GH	Growth hormone	GH	P01241
CVD2_142_FS	Follistatin	FS	P19883
CVD2_143_GLO1	Lactoylglutathione lyase	GLO1	Q04760
CVD2_144_CD84	SLAM family member 5	CD84	Q9UIB8
CVD2_145_PAPPA	Pappalysin-1	PAPPA	Q13219
CVD2_148_SERPINA12	Serpin A12	SERPINA12	Q8IW75
CVD2_149_REN	Renin	REN	P00797
CVD2_150_DECR1	2,4-dienoyl-CoA reductase, mitochondrial	DECR1	Q16698
CVD2_151_MERTK	Tyrosine-protein kinase Mer	MERTK	Q12866
CVD2_152_TIM	Kidney Injury Molecule	KIM1	Q96D42
CVD2_153_THBS2	Thrombospondin-2	THBS2	P35442
CVD2_154_TM	Thrombomodulin	TM	P07204
CVD2_155_VSIG2	V-set and immunoglobulin domain-containing protein 2	VSIG2	Q96IQ7
CVD2_156_AMBP	Protein AMBP	AMBP	P02760
CVD2_157_PRELP	Prolargin	PRELP	P51888
CVD2_158_HO.1	Heme oxygenase 1	HO-1	P09601
CVD2_159_XCL1	Lymphotactin	XCL1	P47992
CVD2_160_IL16	Pro-interleukin-16	IL16	Q14005
CVD2_161_SORT1	Sortilin	SORT1	Q99523
CVD2_162_CEACAM8	Carcinoembryonic antigenrelated cell adhesion molecule 8	CEACAM8	P31997
CVD2_163_PTX3	Pentraxin-related protein PTX3	PTX3	P26022
CVD2_164_PSGL.1	P-selectin glycoprotein ligand 1	PSGL-1	Q14242
CVD2_165_CCL17	C-C motif chemokine 17	CCL17	Q92583
CVD2_166_CCL3	C-C motif chemokine 3	CCL3	P10147
CVD2_167_MMP.7	Matrix metalloproteinase-7	MMP-7	P09237
CVD2_168_IgG.Fc.receptor.II.b	Low affinity immunoglobulin gamma Fc region receptor II-b	IgG Fc receptor II-b	P31994
CVD2_169_ITGB1BP2	Melusin	ITGB1BP2	Q9UKP3
CVD2_170_DCN	Decorin	DCN	P07585
CVD2_171_Dkk.1	Dickkopf-related protein 1	Dkk-1	O94907
CVD2_172_LPL	Lipoprotein lipase	LPL	P06858
CVD2_173_PRSS8	Prostasin	PRSS8	Q16651
CVD2_174_AGRP	Agouti-related protein	AGRP	O00253
CVD2_175_HB.EGF	Proheparin-binding EGF-like growth factor	HB-EGF	Q99075
CVD2_176_GDF.2	Growth/differentiation factor 2	GDF-2	Q9UK05
CVD2_177_FABP2	Fatty acid-binding protein, intestinal	FABP2	P12104
CVD2_178_THPO	Thrombopoietin	THPO	P40225
CVD2_179_MARCO	Macrophage receptor MARCO	MARCO	Q9UEW3

Figure Display Name	Protein Name	Symbol	UniProt ID
CVD2_180_GT	Gastrotropin	GT	P51161
CVD2_181_BNP	Natriuretic peptides B	BNP	P16860
CVD2_182_MMP.12	Matrix metalloproteinase-12	MMP-12	P39900
CVD2_183_ACE2	Angiotensin-converting enzyme 2	ACE2	Q9BYF1
CVD2_184_PD.L2	Programmed cell death 1 ligand 2	PD-L2	Q9BQ51
CVD2_185_CTSL1	Cathepsin L1	CTSL1	P07711
CVD2_186_hOSCAR	Osteoclast-associated immunoglobulin-like receptor	hOSCAR	Q8IYS5
CVD2_187_TNFRSF13B	Tumour necrosis factor receptor superfamily member 13B	TNFRSF13B	O14836
CVD2_188_TGM2	Protein-glutamine gamma-glutamyltransferase 2	TGM2	P21980
CVD2_189_LEP	Leptin	LEP	P41159
CVD2_190_CA5A	Carbonic anhydrase 5A, mitochondrial	CA5A	P35218
CVD2_191_HSP.27	Heat shock 27 kDa protein	HSP 27	P04792
CVD2_192_CD4	T-cell surface glycoprotein CD4	CD4	P01730
CVD2_193_NEMO	NF-kappa-B essential modulator	NEMO	Q9Y6K9
CVD2_194_VEGF.D	Vascular endothelial growth factor D	VEGFD	O43915
CVD2_195_PARP.1	Poly [ADP-ribose] polymerase 1	PARP-1	P09874
CVD2_196_HAOX1	Hydroxyacid oxidase 1	HAOX1	Q9UJM8
CVD3_101_TNFRSF14	Tumour necrosis factor receptor superfamily member 14	TNFRSF14	Q92956
CVD3_102_LDL.receptor	Low-density lipoprotein receptor	LDL receptor	P01130
CVD3_103_ITGB2	Integrin beta-2	ITGB2	P05107
CVD3_105_IL.17RA	Interleukin-17 receptor A	IL-17RA	Q96F46
CVD3_106_TNF.R2	Tumour necrosis factor receptor 2	TNF-R2	P20333
CVD3_107_MMP.9	Matrix metalloproteinase-9	MMP-9	P14780
CVD3_108_EPHB4	Ephrin type-B receptor 4	EPHB4	P54760
CVD3_109_IL2.RA	Interleukin-2 receptor subunit alpha	IL2-RA	P01589
CVD3_110_OPG	Osteoprotegerin	OPG	O00300
CVD3_111_ALCAM	CD166 antigen	ALCAM	Q13740
CVD3_112_TFF3	Trefoil factor 3	TFF3	Q07654
CVD3_113_SELP	P-selectin	SELP	P16109
CVD3_114_CSTB	Cystatin-B	CSTB	P04080
CVD3_115_MCP.1	Monocyte chemotactic protein 1	MCP-1	P13500
CVD3_116_CD163	Scavenger receptor cysteine-rich type 1 protein M130	CD163	Q86VB7
CVD3_117_Gal.3	Galectin-3	Gal-3	P17931
CVD3_118_GRN	Granulins	GRN	P28799
CVD3_120_MEPE	Matrix extracellular phosphoglycoprotein	MEPE	Q9NQ76
CVD3_121_BLM.hydrolase	Bleomycin hydrolase	BLM hydrolase	Q13867
CVD3_122_PLC	Perlecan	PLC	P98160
CVD3_123_LTBR	Lymphotoxin-beta receptor	LTBR	P36941
CVD3_124_Notch.3	Neurogenic locus notch homolog protein 3	Notch 3	Q9UM47
CVD3_125_TIMP4	Metalloproteinase inhibitor 4	TIMP4	Q99727
CVD3_126_CNTN1	Contactin-1	CNTN1	Q12860
CVD3_127_CDH5	Cadherin-5	CDH5	P33151
CVD3_128_TLT.2	Trem-like transcript 2 protein	TLT-2	Q5T2D2
CVD3_129_FABP4	Fatty acid-binding protein, adipocyte	FABP4	P15090
CVD3_130_TFPI	Tissue factor pathway inhibitor	TFPI	P10646
CVD3_131_PAI	Plasminogen activator inhibitor 1	PAI	P05121
CVD3_132_CCL24	C-C motif chemokine 24	CCL24	O00175
CVD3_133_TR	Transferrin receptor protein 1	TR	P02786
CVD3_134_TNFRSF10C	Tumour necrosis factor receptor superfamily member 10C	TNFRSF10C	O14798
CVD3_135_GDF.15	Growth/differentiation factor 15	GDF-15	Q99988
CVD3_136_SELE	E-selectin	SELE	P16581
CVD3_137_AZU1	Azurocidin	AZU1	P20160
CVD3_138_DLK.1	Protein delta homolog 1	DLK-1	P80370
CVD3_139_SPON1	Spondin-1	SPON1	Q9HCB6
CVD3_140_MPO	Myeloperoxidase	MPO	P05164
CVD3_141_CXCL16	C-X-C motif chemokine 16	CXCL16	Q9H2A7
CVD3_142_IL.6RA	Interleukin-6 receptor subunit alpha	IL-6RA	P08887
CVD3_143_RETN	Resistin	RETN	Q9HD89
CVD3_144_IGFBP.1	Insulin-like growth factor-binding protein 1	IGFBP-1	P08833
CVD3_145_CHIT1	Chitotriosidase-1	CHIT1	Q13231
CVD3_148_TR.AP	Tartrate-resistant acid phosphatase type 5	TR-AP	P13686
CVD3_150_PSP.D	Pulmonary surfactant-associated protein D	PSP-D	P35247
CVD3_151_PI3	Elafin	PI3	P19957
CVD3_152_Ep.CAM	Epithelial cell adhesion molecule	Ep-CAM	P16422
CVD3_153_AP.N	Aminopeptidase N	AP-N	P15144
CVD3_154_AXL	Tyrosine-protein kinase receptor UFO	AXL	P30530
CVD3_155_IL.1RT1	Interleukin-1 receptor type 1	IL-1RT1	P14778
CVD3_156_MMP.2	Matrix metalloproteinase-2	MMP-2	P08253
CVD3_157_FAS	Tumour necrosis factor receptor superfamily member 6	FAS	P25445
CVD3_158_MB	Myoglobin	MB	P02144
CVD3_159_TNFSF13B	Tumour necrosis factor ligand superfamily member 13B	TNFSF13B	Q9Y275
CVD3_160_PRTN3	Myeloblastin	PRTN3	P24158
CVD3_161_PCSK9	Proprotein convertase subtilisin/kexin type 9	PCSK9	Q8NBP7
CVD3_162_U.PAR	Urokinase plasminogen activator surface receptor	U-PAR	Q03405
CVD3_163_OPN	Osteopontin	OPN	P10451
CVD3_164_CTSD	Cathepsin D	CTSD	P07339
CVD3_165_PGLYRP1	Peptidoglycan recognition protein 1	PGLYRP1	O75594
CVD3_166_CPA1	Carboxypeptidase A1	CPA1	P15085
CVD3_167_JAM.A	Junctional adhesion molecule A	JAM-A	Q9Y624
CVD3_168_Gal.4	Galectin-4	Gal-4	P56470
CVD3_169_IL.1RT2	Interleukin-1 receptor type 2	IL-1RT2	P27930
CVD3_170_SHPS.1	Tyrosine-protein phosphatase non-receptor type substrate 1	SHPS-1	P78324
CVD3_171_CCL15	C-C motif chemokine 15	CCL15	Q16663
CVD3_172_CASP.3	Caspase-3	CASP-3	P42574
CVD3_173_uPA	Urokinase-type plasminogen activator	uPA	P00749

Figure Display Name	Protein Name	Symbol	UniProt ID
CVD3_174_CPB1	Carboxypeptidase B	CPB1	P15086
CVD3_175_CHI3L1	Chitinase-3-like protein 1	CHI3L1	P36222
CVD3_176_ST2	ST2 protein	ST2	Q01638
CVD3_177_t.PA	Tissue-type plasminogen activator	t-PA	P00750
CVD3_178_SCGB3A2	Secretoglobin family 3A member 2	SCGB3A2	Q96PL1
CVD3_179_EGFR	Epidermal growth factor receptor	EGFR	P00533
CVD3_180_IGFBP.7	Insulin-like growth factor-binding protein 7	IGFBP-7	Q16270
CVD3_181_CD93	Complement component C1q receptor	CD93	Q9NPY3
CVD3_182_IL.18BP	Interleukin-18-binding protein	IL-18BP	O95998
CVD3_183_COL1A1	Collagen alpha-1(I) chain	COL1A1	P02452
CVD3_184_PON3	Paraoxonase	PON3	Q15166
CVD3_185_CTSZ	Cathepsin Z	CTSZ	Q9UBR2
CVD3_186_MMP.3	Matrix metalloproteinase-3	MMP-3	P08254
CVD3_187_RARRES2	Retinoic acid receptor responder protein 2	RARRES2	Q99969
CVD3_188_ICAM.2	Intercellular adhesion molecule 2	ICAM-2	P13598
CVD3_189_KLK6	Kallikrein-6	KLK6	Q92876
CVD3_190_PDGF.subunit.A	Platelet-derived growth factor subunit A	PDGF subunit A	P04085
CVD3_191_TNF.R1	Tumour necrosis factor receptor 1	TNF-R1	P19438
CVD3_192_IGFBP.2	Insulin-like growth factor-binding protein 2	IGFBP-2	P18065
CVD3_193_vWF	von Willebrand factor	vWF	P04275
CVD3_194_PECAM.1	Platelet endothelial cell adhesion molecule	PECAM-1	P16284
CVD3_195_NT.pro.BNP	N-terminal prohormone brain natriuretic peptide	NT-proBNP	NA
CVD3_196_CCL16	C-C motif chemokine 16	CCL16	O15467
INF_101_IL.8	Interleukin-8	IL-8	P10145
INF_102_VEGF.A	Vascular endothelial growth factor A	VEGF-A	P15692
INF_105_MCP.3	Monocyte chemotactic protein 3	MCP-3	P80098
INF_106_hGDNF	Glial cell line-derived neurotrophic factor	GDNF	P39905
INF_107_CDCP1	CUB domain-containing protein 1	CDCP1	Q9H5V8
INF_108_CD244	Natural killer cell receptor 2B4	CD244	Q9BZW8
INF_109_IL.7	Interleukin-7	IL-7	P13232
INF_111_LAP.TGF.beta.1	Latency-associated peptide transforming growth factor beta-1	Lap TGF beta 1	P01137
INF_114_IL.17C	Interleukin-17C	IL-17C	Q9P0M4
INF_116_IL.17A	Interleukin-17A	IL-17A	Q16552
INF_117_CXCL11	C-X-C motif chemokine 11	CXCL11	O14625
INF_118_AXIN1	Axin-1	AXIN1	O15169
INF_120_TRAIL	TNF-related apoptosis-inducing ligand	TRAIL	P50591
INF_121_IL.20RA	Interleukin-20 receptor subunit alpha	IL-20RA	Q9UHF4
INF_122_CXCL9	C-X-C motif chemokine 9	CXCL9	Q07325
INF_123_CST5	Cystatin D	CST5	P28325
INF_124_IL.2RB	Interleukin-2 receptor subunit alpha	IL2-RA	P01589
INF_125_IL.1.alpha	Interleukin-2 receptor subunit beta	IL-2RB	P14784
INF_126_OSM	Oncostatin-M	OSM	P13725
INF_127_IL.2	N-terminal prohormone brain natriuretic peptide	NT-proBNP	NA
INF_129_TSLP	Thymic stromal lymphopoietin	TSLP	Q969D9
INF_130_CCL4	C-C motif chemokine 4	CCL4	P13236
INF_131_CD6	T cell surface glycoprotein CD6 isoform	CD6	Q8WWJ7
INF_134_SLAMF1	Signaling lymphocytic activation molecule	SLAMF1	Q13291
INF_135_TGFA	Transforming growth factor alpha	TGF-alpha	P01135
INF_136_MCP.4	Monocyte chemotactic protein 4	MCP-4	Q99616
INF_137_CCL11	Eotaxin	CCL11	P51671
INF_138_TNFSF14	Tumour necrosis factor ligand superfamily member 14	TNFSF14	O43557
INF_140_IL.10RA	Interleukin-10 receptor subunit alpha	IL-10RA	Q13651
INF_141_FGF.5	Fibroblast growth factor 5	FGF-5	Q8NF90
INF_142_MMP.1	Matrix metalloproteinase-1	MMP-1	P03956
INF_143_LIF.R	Leukemia inhibitory factor receptor	LIF-R	P42702
INF_145_CCL19	C-C motif chemokine 19	CCL19	Q99731
INF_148_IL.15RA	Interleukin-15 receptor subunit alpha	IL-15RA	Q13261
INF_149_IL.10RB	Interleukin-10 receptor subunit beta	IL-10RB	Q08334
INF_150_IL.22.RA1	Interleukin-22 receptor subunit alpha-1	IL-22 RA1	Q8N6P7
INF_151_IL.18R1	Interleukin-18 receptor 1	IL-18R1	Q13478
INF_152_PD.L1	Programmed cell death 1 ligand 1	PD-L1	Q9NZQ7
INF_153_Beta.NGF	Beta-nerve growth factor	Beta-NGF	P01138
INF_154_CXCL5	C-X-C motif chemokine 5	CXCL5	P42830
INF_155_TRANCE	TNF-related activation-induced cytokine	TRANCE	O14788
INF_156_HGF	Hepatocyte growth factor	HGF	P14210
INF_157_IL.12B	Interleukin-12 subunit beta	IL-12B	P29460
INF_158_IL.24	Interleukin-24	IL-24	Q13007
INF_159_IL.13	Interleukin-13	IL-13	P35225
INF_160_ARTN	Artemin	ARTN	Q5T4W7
INF_161_MMP.10	Matrix metalloproteinase-10	MMP-10	P09238
INF_162_IL.10	Interleukin-10	IL10	P22301
INF_163_TNF	Tumour necrosis factor	TNF	P01375
INF_164_CCL23	C-C motif chemokine 23	CCL23	P55773
INF_165_CD5	T-cell surface glycoprotein CD5	CD5	P06127
INF_167_Flt3L	Fms-related tyrosine kinase 3 ligand	Flt3L	P49771
INF_168_CXCL6	C-X-C motif chemokine 6	CXCL6	P80162
INF_169_CXCL10	C-X-C motif chemokine 10	CXCL10	P02778
INF_170_4E.BP1	Eukaryotic translation initiation factor 4E-binding protein 1	4E-BP1	Q13541
INF_171_IL.20	Interleukin-20	IL-20	Q9NYY1
INF_172_SIRT2	SIR2-like protein 2	SIRT2	Q8IXJ6
INF_173_CCL28	C-C motif chemokine 28	CCL28	Q9NRJ3
INF_174_DNER	Delta and Notch-like epidermal growth factor-related receptor	DNER	Q8NFT8
INF_175_EN.RAGE	Protein S100-A12	EN-RAGE	P80511
INF_176_CD40	CD40L receptor	CD40	P25942
INF_177_IL.33	Interleukin-33	IL-33	O95760
INF_178_IFN.gamma	Interferon gamma	IFN-gamma	P01579

Figure Display Name	Protein Name	Symbol	UniProt ID
INF_179_FGF.19	Fibroblast growth factor 19	FGF-19	O95750
INF_180_IL.4	Interleukin-4	IL-4	P05112
INF_181_LIF	Leukemia inhibitory factor	LIF	P15018
INF_182_NRTN	Neurturin	NRTN	Q99748
INF_183_MCP.2	Monocyte chemotactic protein 2	MCP-2	P80075
INF_184_CASP.8	Caspase-8	CASP-8	Q14790
INF_185_CCL25	C-C motif chemokine 25	CCL25	O15444
INF_186_CX3CL1	Fractalkine	CX3CL1	P78423
INF_187_TNFRSF9	Tumour necrosis factor receptor superfamily member 9	TNFRSF9	Q07011
INF_188_NT.3	Neurotrophin-3	NT-3	P20783
INF_189_TWEAK	Tumour necrosis factor (Ligand) superfamily, member 12	TWEAK	O43508
INF_190_CCL20	C-C motif chemokine 20	CCL20	P78556
INF_191_ST1A1	Sulfotransferase 1A1	ST1A1	P50225
INF_192_STAMPB	STAM-binding protein	STAMBP	O95630
INF_193_IL.5	Interleukin-5	IL5	P05113
INF_194_ADA	Adenosine Deaminase	ADA	P00813
INF_195_TNFB	TNF-beta	TNFB	P01374
INF_196_CSF.1	Macrophage colony-stimulating factor 1	CSF-1	P09603
CVD1_AM	Adrenomedullin	AM	NA
CVD1_CA125	Ovarian cancer-related tumour marker 125	CA125	NA
CVD1_ECP	Eosinophil cationic protein	ECP	NA
CVD1_EGF	Epidermal growth factor	EGF	NA
CVD1_ESM1	Endothelial cell-specific molecule 1	ESM1	NA
CVD1_GAL	Galanin peptides	GAL	NA
CVD1_hK11	Kallikrein-11	hK11	NA
CVD1_PRL	Prolactin	PRL	NA
Non-Olink Phenotypes			
Display Name	**Phenotype**		
alc_gpW	Alcohol consumption	-	-
bmi	Body Mass Index	-	-
eGFR	Estimated Glomerular Filtration Rate	-	-

Supplementary Table 2 – Serum urate-Olink lasso regression mean coefficients

Phenotype	Cohort	Count	Coef. Mean	Coef. SD
BMI	Vis & ORCADES	1000	0.073	0.002
	EGCUT	1000	0.100	0.002
	INTERVAL	7	0.000	0.000
	Lifelines DEEP	1000	0.090	0.001
FGF-21	Vis & ORCADES	1000	0.077	0.002
	EGCUT	1000	0.041	0.007
	INTERVAL	600	0.010	0.010
	Lifelines DEEP	NA	NA	NA
CCL3	Vis & ORCADES	1000	0.010	0.000
	EGCUT	1000	0.017	0.002
	INTERVAL	0	-	-
	Lifelines DEEP	NA	NA	NA
LDL-receptor	Vis & ORCADES	999	0.020	0.005
	EGCUT	0	-	-
	INTERVAL	0	-	-
	Lifelines DEEP	1000	0.058	0.009
PLC	Vis & ORCADES	1000	0.049	0.013
	EGCUT	1000	0.029	0.005
	INTERVAL	0	-	-
	Lifelines DEEP	0	-	-
FABP4	Vis & ORCADES	1000	0.101	0.005
	EGCUT	1000	0.085	0.010
	INTERVAL	1000	0.048	0.006
	Lifelines DEEP	1000	0.020	0.004
MMP-2	Vis & ORCADES	1000	- 0.042	0.011
	EGCUT	138	- 0.001	0.002
	INTERVAL	0	-	-
	Lifelines DEEP	892	- 0.012	0.011
CHI3L1	Vis & ORCADES	1000	0.030	0.002
	EGCUT	598	0.007	0.008

Phenotype	Cohort	Count	Coef. Mean	Coef. SD
	INTERVAL	0	-	-
	Lifelines DEEP	0	-	-
PON3	Vis & ORCADES	1000	- 0.070	0.001
	EGCUT	1000	- 0.026	0.004
	INTERVAL	0	-	-
	Lifelines DEEP	1000	- 0.025	0.004
IGFBP-2	Vis & ORCADES	1000	- 0.137	0.005
	EGCUT	851	- 0.015	0.010
	INTERVAL	7	- 0.000	0.000
	Lifelines DEEP	1000	- 0.074	0.005
eGFR	Vis & ORCADES	1000	- 0.106	0.001
	EGCUT	1000	- 0.126	0.016
	INTERVAL	1000	- 0.107	0.019
	Lifelines DEEP	1000	- 0.073	0.015

Supplementary Table 3 - Serum urate-lipidomic lasso regression mean coefficients.

For all phenotypes appearing in 95% of models in one or more analyses.

Phenotype	Subset	Count	Coef. Mean	Coef. SD
alc_gpW	All	581	0.006	0.007
alc_gpW	CKD excluded	984	0.015	0.004
alc_gpW	Female	2	0.000	0.000
alc_gpW	Male	0	0.000	0.000
bmi	All	1000	0.195	0.003
bmi	CKD excluded	1000	0.197	0.003
bmi	Female	1000	0.197	0.006
bmi	Male	1000	0.128	0.009
C_18__0	All	986	0.027	0.009
C_18__0	CKD excluded	1000	0.026	0.004
C_18__0	Female	938	0.008	0.003
C_18__0	Male	0	0.000	0.000
eGFR	All	1000	-0.157	0.023
eGFR	CKD excluded	1000	-0.118	0.014
eGFR	Female	1000	-0.113	0.017
eGFR	Male	753	-0.014	0.015
LPC_15__0	All	80	0.000	0.001
LPC_15__0	CKD excluded	0	0.000	0.000
LPC_15__0	Female	974	-0.014	0.007
LPC_15__0	Male	0	0.000	0.000
LPC_20__0	All	1000	-0.063	0.002
LPC_20__0	CKD excluded	1000	-0.055	0.001
LPC_20__0	Female	1000	-0.038	0.005
LPC_20__0	Male	884	-0.031	0.019
LPC_22__4	All	958	-0.034	0.022
LPC_22__4	CKD excluded	999	-0.053	0.014
LPC_22__4	Female	790	-0.021	0.019
LPC_22__4	Male	2	0.000	0.000
PC_30__1	All	801	-0.011	0.009
PC_30__1	CKD excluded	984	-0.019	0.006
PC_30__1	Female	945	-0.023	0.011
PC_30__1	Male	0	0.000	0.000
PC_38__3	All	801	0.011	0.009
PC_38__3	CKD excluded	1000	0.027	0.004
PC_38__3	Female	783	0.004	0.002
PC_38__3	Male	0	0.000	0.000
PC_O_34__2	All	986	-0.017	0.009
PC_O_34__2	CKD excluded	960	-0.011	0.004
PC_O_34__2	Female	0	0.000	0.000
PC_O_34__2	Male	84	-0.001	0.003
PC_O_36__2	All	1000	-0.060	0.008
PC_O_36__2	CKD excluded	1000	-0.055	0.008
PC_O_36__2	Female	1000	-0.060	0.003
PC_O_36__2	Male	0	0.000	0.000
PC_O_40__6	All	404	-0.007	0.010
PC_O_40__6	CKD excluded	936	-0.030	0.012
PC_O_40__6	Female	994	-0.021	0.005
PC_O_40__6	Male	0	0.000	0.000
PE_32__1	All	1000	0.033	0.008
PE_32__1	CKD excluded	1000	0.029	0.006
PE_32__1	Female	1000	0.047	0.008
PE_32__1	Male	0	0.000	0.000
PE_40__6	All	1000	0.183	0.014
PE_40__6	CKD excluded	1000	0.218	0.020
PE_40__6	Female	1000	0.121	0.010
PE_40__6	Male	1000	0.183	0.017
PLPE_16__0_20__5	All	581	0.011	0.012

Phenotype	Subset	Count	Coef. Mean	Coef. SD
PLPE_16__0_20__5	CKD excluded	960	0.014	0.005
PLPE_16__0_20__5	Female	0	0.000	0.000
PLPE_16__0_20__5	Male	2	0.000	0.000
sbp	All	998	0.030	0.008
sbp	CKD excluded	999	0.022	0.003
sbp	Female	994	0.031	0.011
sbp	Male	0	0.000	0.000
SPM_15__0	All	494	-0.007	0.009
SPM_15__0	CKD excluded	984	-0.024	0.007
SPM_15__0	Female	994	-0.028	0.009
SPM_15__0	Male	0	0.000	0.000
SPM_dih_18__0	All	494	0.005	0.006
SPM_dih_18__0	CKD excluded	997	0.029	0.007
SPM_dih_18__0	Female	90	0.001	0.002
SPM_dih_18__0	Male	0	0.000	0.000

Supplementary Table 4 – 183 index SNPs identified in the CKDGen trans-ethnic meta-analysis

RSID	Chr.	position (b37)	A1/A2	EAF	Function	Gene Name	Effect (mg/dL)	SE	p-value	I2 (%)	P_{het}	$P_{anc-het}$	Gout OR	Gout OR 95% CI	Gout OR 95% CI
rs10803394	1	15,909,480	C/G	0.373	intron	*AGMAT*	-0.020	0.0032	5.83E-10	0	6.92E-01	3.50E-01	0.96	0.93	0.98
rs79598313	1	27,284,913	T/C	0.026	intron	*KDF1*	0.102	0.0127	1.16E-15	10.7	2.43E-01	8.00E-02	1.39	1.28	1.50
rs10890263	1	44,061,032	T/C	0.737	intron	*PTPRF*	-0.020	0.0037	4.83E-08	17.9	1.01E-01	7.50E-01	0.99	0.96	1.02
rs584425	1	48,536,798	A/G	0.333	unknown	*SKINT1L*	-0.021	0.0033	1.36E-10	0.9	4.56E-01	6.20E-01	0.98	0.95	1.00
rs2356864	1	50,839,740	A/G	0.568	unknown	*DMRTA2*	0.020	0.0034	1.69E-09	0	8.96E-01	5.30E-01	1.02	1.00	1.05
rs662026	1	91,531,022	A/G	0.800	unknown	*ZNF644*	0.021	0.0038	3.26E-08	0	7.25E-01	8.30E-01	1.04	1.00	1.07
rs7417952	1	93,854,186	C/G	0.396	unknown	*DR1*	-0.023	0.0037	2.75E-10	0	5.51E-01	9.00E-01	0.95	0.90	1.00
rs141990161	1	119,943,525	T/C	0.985	unknown	*HAO2*	0.133	0.0235	1.59E-08	26.4	7.11E-02	8.40E-02	1.19	1.04	1.36
rs10910845	1	145,723,120	A/C	0.539	unknown	*NBPF20*	0.060	0.0034	1.10E-69	0	6.63E-01	4.40E-01	1.12	1.09	1.15
rs11204682	1	150,595,537	T/G	0.193	intron, near-gene-3'	*ENSA*	0.032	0.0043	6.00E-14	0.9	4.55E-01	4.00E-02	1.03	1.00	1.06
rs2070803	1	155,157,715	A/G	0.497	near-gene-3'	*TRIM46*	0.054	0.0033	2.12E-60	14	1.61E-01	8.20E-01	1.08	1.05	1.11
rs12134456	1	155,722,506	C/G	0.635	intron	*GON4L*	-0.044	0.0048	6.33E-20	4.4	3.76E-01	7.10E-01	0.94	0.92	0.97
rs2760215	1	163,675,883	T/C	0.451	unknown	*LOC100422212*	-0.022	0.0033	7.55E-12	0	7.41E-01	4.60E-01	0.97	0.95	1.00
rs56129505	1	186,718,261	T/C	0.288	unknown	*PACERR*	0.021	0.0037	2.61E-08	0	5.65E-01	8.40E-01	1.03	1.00	1.06
rs2970581	1	212,083,512	A/G	0.036	unknown	*INTS7*	0.059	0.0100	3.78E-09	0	6.15E-01	3.60E-03	1.12	1.04	1.19
rs6746275	2	635,857	A/C	0.847	unknown	*TMEM18*	0.035	0.0043	5.50E-16	14.8	1.47E-01	6.50E-01	1.04	1.00	1.07
rs10177191	2	9,246,721	T/C	0.400	unknown	*ASAP2*	-0.018	0.0032	3.05E-08	0	8.85E-01	7.50E-01	0.98	0.95	1.00
rs72782806	2	15,788,511	A/G	0.230	unknown	*DDX1*	0.023	0.0038	1.02E-09	0	5.32E-01	6.70E-01	1.02	0.99	1.05
rs72804857	2	27,161,476	C/G	0.150	intron	*DPYSL5*	0.028	0.0044	2.53E-10	16.3	1.24E-01	4.50E-01	1.06	1.02	1.10
rs1260326	2	27,730,940	T/C	0.436	missense	*GCKR*	0.066	0.0032	2.97E-95	44.3	6.22E-05	1.70E-03	1.21	1.18	1.24
rs62140395	2	28,244,926	C/G	0.117	intron	*BABAM2*	0.050	0.0064	8.68E-15	6.4	3.28E-01	4.70E-01	1.11	1.07	1.16
rs10084334	2	37,250,891	C/G	0.617	intron	*HEATR5B*	0.019	0.0034	4.33E-08	0	7.31E-01	1.70E-01	1.00	0.97	1.03
rs142874192	2	54,900,899	C/G	0.957	unknown	*SPTBN1*	-0.059	0.0103	1.08E-08	0	9.52E-01	8.90E-01	0.94	0.83	1.06
rs6730325	2	59,315,828	A/G	0.667	unknown	*LINC01122*	-0.019	0.0034	2.58E-08	0	5.79E-01	4.60E-01	0.97	0.95	1.00
rs7572603	2	61,552,145	C/G	0.639	intron	*USP34*	-0.023	0.0033	3.62E-12	0	8.43E-01	4.00E-01	0.95	0.92	0.97
rs12987661	2	69,813,458	T/C	0.879	intron	*AAK1*	0.043	0.0052	4.32E-16	0	8.61E-01	8.30E-01	1.05	1.01	1.09
rs759219	2	71,163,225	T/C	0.413	intron	*ATP6V1B1*	-0.018	0.0032	9.23E-09	0	6.77E-01	3.80E-02	0.99	0.96	1.01
rs6707386	2	113,981,022	A/G	0.354	intron	*PAX8*	0.018	0.0032	3.93E-08	0	9.93E-01	5.70E-01	1.02	1.00	1.05
rs17050272	2	121,306,440	A/G	0.430	unknown	*LINC01101*	0.028	0.0032	1.01E-17	22.4	5.13E-02	9.00E-02	1.02	0.99	1.05
rs2304667	2	121,989,489	A/G	0.421	coding-syn. intron	*TFCP2L1*	0.023	0.0038	1.02E-09	0	8.89E-01	6.60E-01	1.02	0.97	1.08
rs11683692	2	145,509,615	T/C	0.946	intron	*TEX41*	-0.047	0.0083	1.55E-08	10.5	2.37E-01	8.50E-01	1.00	0.95	1.06
rs2307394	2	148,716,428	T/C	0.632	missense, untranslated-5'	*ORC4*	-0.020	0.0032	1.05E-09	6.9	3.15E-01	4.20E-01	0.97	0.94	1.00
rs113704612	2	158,294,018	T/G	0.035	intron	*CYTIP*	-0.080	0.0118	1.14E-11	0	5.93E-01	1.90E-01	0.91	0.83	1.00
rs1457231	2	161,101,562	T/C	0.705	intron	*LINC02478*	0.019	0.0035	3.17E-08	0	9.34E-01	3.10E-01	1.02	0.99	1.05
rs2075251	2	170,011,458	A/T	0.657	intron	*LRP2*	-0.037	0.0035	7.01E-27	1.5	4.42E-01	1.80E-01	0.92	0.89	0.95
rs10198459	2	177,273,272	T/C	0.253	unknown	*MTX2*	0.027	0.0036	1.27E-13	0	6.90E-01	3.20E-01	1.06	1.03	1.09
rs3769810	2	183,037,246	A/G	0.757	intron	*PDE1A*	0.022	0.0037	2.81E-09	0	8.67E-01	1.70E-01	1.05	1.02	1.09
rs1047891	2	211,540,507	A/C	0.287	missense	*CPS1*	-0.024	0.0036	1.91E-11	1.4	4.44E-01	1.90E-02	0.97	0.94	1.00
rs1949651	2	213,117,629	T/C	0.519	intron	*ERBB4*	0.021	0.0032	2.64E-11	7.4	2.99E-01	7.60E-01	1.04	1.01	1.07
rs73058028	3	31,528,128	C/G	0.644	unknown	*STT3B*	0.018	0.0032	1.54E-08	0	5.69E-01	4.60E-01	1.00	0.97	1.02
rs699465	3	52,310,442	A/G	0.148	intron	*WDR82*	0.033	0.0046	4.60E-13	0	6.30E-01	2.80E-03	1.08	1.04	1.12
rs2244552	3	53,055,522	A/G	0.547	intron	*SFMBT1*	-0.043	0.0031	1.06E-43	10.2	2.37E-01	2.10E-02	0.91	0.89	0.94
rs7039	3	69,154,343	C/G	0.446	untranslated-3'	*ARL6IP5*	-0.021	0.0031	5.24E-11	6.1	3.32E-01	6.00E-01	0.97	0.94	0.99
rs9859616	3	125,149,488	A/G	0.219	unknown	*SNX4*	-0.024	0.0038	4.17E-10	0	8.88E-01	3.80E-01	0.98	0.95	1.01

RSID	Chr.	position (b37)	A1/A2	EAF	Function	Gene Name	Effect (mg/dL)	SE	p-value	I2 (%)	P_{het}	$P_{anc-het}$	Gout OR	Gout OR 95% CI	Gout OR 95% CI
rs11718633	3	126,012,421	T/C	0.194	unknown	*KLF15*	-0.027	0.0046	7.64E-09	0	6.15E-01	5.80E-01	0.96	0.93	0.99
rs78946096	3	132,188,163	A/G	0.947	intron	*DNAJC13*	-0.060	0.0099	1.48E-09	8	3.08E-01	9.70E-01	0.91	0.86	0.96
rs12496412	3	141,741,823	A/G	0.668	intron	*TFDP2*	0.021	0.0033	5.30E-10	0	9.93E-01	4.60E-01	1.01	0.99	1.04
rs6805417	3	142,762,487	T/C	0.343	intron	*U2SURP*	0.019	0.0033	1.10E-08	0	9.21E-01	3.80E-01	1.03	1.01	1.06
rs60388273	3	149,214,543	A/G	0.228	intron	*TM4SF4*	0.030	0.0044	3.15E-11	3.7	3.88E-01	3.00E-01	1.03	0.98	1.08
rs62294340	3	169,155,476	A/G	0.354	intron	*MECOM*	-0.021	0.0032	7.62E-11	0	6.93E-01	5.90E-01	0.96	0.93	0.99
rs1533096	4	4,784,139	C/G	0.595	intron	*LOC101928279*	0.020	0.0036	3.16E-08	0	5.43E-01	6.40E-01	0.99	0.96	1.01
rs6820627	4	9,491,205	A/G	0.071	unknown	*DEFB131A*	-0.100	0.0107	4.66E-21	28.6	2.54E-02	3.20E-05	0.89	0.84	0.95
rs3775947	4	9,995,240	T/C	0.691	intron	*SLC2A9*	0.277	0.0033	0.00E+00	90	1.10E-109	1.50E-127	1.51	1.47	1.56
rs12504795	4	10,499,344	T/C	0.741	intron	*CLNK*	0.075	0.0035	4.25E-101	51.4	2.74E-07	1.70E-18	1.15	1.11	1.18
rs11940694	4	39,414,993	A/G	0.464	intron	*KLB*	-0.017	0.0031	2.37E-08	0	8.75E-01	9.90E-02	0.95	0.93	0.98
rs10857147	4	81,181,072	A/T	0.712	unknown	*FGF5*	0.028	0.0035	1.67E-15	0	9.73E-01	3.00E-01	1.04	1.01	1.07
rs116183010	4	88,468,158	A/G	0.024	unknown	*SPARCL1*	0.096	0.0138	3.66E-12	22.7	6.67E-02	2.10E-01	1.29	1.19	1.41
rs74904971	4	89,050,026	A/C	0.196	intron	*ABCG2*	0.217	0.0043	0.00E+00	73.7	1.58E-25	3.90E-25	2.04	1.96	2.12
rs35232147	4	89,916,224	T/C	0.373	intron	*FAM13A*	0.018	0.0032	3.37E-08	0	4.85E-01	1.70E-01	1.05	1.03	1.08
rs11097693	4	101,121,391	A/G	0.584	intron	*LOC101929353*	-0.019	0.0031	9.03E-10	0	7.48E-01	7.70E-01	0.96	0.93	0.98
rs12644329	4	143,634,746	A/G	0.643	intron	*INPP4B*	-0.019	0.0034	8.47E-09	0	5.89E-01	2.90E-01	0.98	0.95	1.00
rs1440411	4	144,158,285	T/C	0.605	unknown	*USP38*	-0.027	0.0033	1.18E-15	0	6.58E-01	6.90E-01	0.98	0.96	1.01
rs455213	5	34,660,235	T/C	0.596	intron	*RAI14*	-0.023	0.0033	2.07E-12	22.3	5.13E-02	2.10E-01	0.96	0.93	0.98
rs116379131	5	39,961,618	A/T	0.034	unknown	*LINC00603*	-0.058	0.0105	3.15E-08	10.4	2.40E-01	2.70E-01	1.01	0.94	1.08
rs10942549	5	72,426,137	C/G	0.301	intron	*TMEM171*	-0.040	0.0035	1.09E-30	26.9	2.22E-02	3.00E-02	0.92	0.90	0.95
rs28362590	5	176,731,452	T/G	0.656	intron, near-gene-5'	*PRELID1*	0.021	0.0034	1.05E-09	0	8.29E-01	1.80E-01	1.05	1.02	1.08
rs7757144	6	1,997,865	A/G	0.626	intron	*GMDS*	-0.022	0.0033	2.31E-11	0	8.14E-01	3.00E-01	0.96	0.94	0.99
rs3904600	6	7,109,665	C/G	0.405	intron	*RREB1*	0.057	0.0038	5.23E-51	0	8.52E-01	2.20E-01	1.08	1.05	1.11
rs77951490	6	25,236,645	A/G	0.053	unknown	*LOC101928663*	0.060	0.0086	2.59E-12	24.1	4.30E-02	5.70E-01	1.08	1.03	1.15
rs1359232	6	25,809,716	A/C	0.401	intron	*SLC17A1*	-0.089	0.0033	5.59E-159	0	9.40E-01	2.20E-01	0.85	0.83	0.87
rs35942569	6	26,339,131	A/G	0.905	unknown	*BTN3A2*	-0.061	0.0068	1.19E-19	16.6	1.26E-01	9.50E-02	0.88	0.84	0.92
rs57440165	6	26,843,517	A/C	0.921	intron	*GUSBP2*	-0.060	0.0077	4.25E-15	13.3	1.92E-01	8.50E-01	0.89	0.85	0.93
rs35501037	6	27,739,566	A/T	0.088	unknown	*LOC100131289*	0.052	0.0069	3.36E-14	0	6.93E-01	9.30E-01	1.11	1.06	1.16
rs7752448	6	28,301,099	A/G	0.881	intron	*ZSCAN31*	-0.040	0.0052	7.44E-15	23.1	4.45E-02	1.80E-02	0.91	0.87	0.94
rs3118365	6	28,839,908	A/G	0.092	unknown	*LINC01623*	0.046	0.0067	5.81E-12	4.7	3.68E-01	6.90E-01	1.11	1.06	1.16
rs429479	6	29,372,323	A/G	0.904	unknown	*OR12D2*	-0.047	0.0067	2.02E-12	0	7.08E-01	1.30E-01	0.90	0.86	0.94
rs753725	6	30,890,871	T/C	0.406	intron	*VARS2*	0.021	0.0034	1.05E-09	3.3	4.04E-01	3.90E-01	1.02	0.99	1.05
rs9271585	6	32,590,879	A/C	0.333	unknown	*HLA-DQA1*	-0.025	0.0036	6.94E-12	15	1.65E-01	3.20E-02	0.98	0.95	1.01
rs742493	6	40,998,167	T/C	0.894	missense	*UNC5CL*	0.035	0.0054	8.26E-11	5.6	3.44E-01	4.00E-01	1.03	0.98	1.07
rs56401710	6	43,269,180	A/C	0.587	intron	*SLC22A7*	-0.042	0.0037	5.00E-30	23.2	1.01E-01	5.50E-01	0.98	0.92	1.04
rs10223666	6	43,805,502	C/G	0.733	unknown	*VEGFA*	0.044	0.0037	1.27E-32	0	9.60E-01	1.30E-01	1.05	1.02	1.08
rs9373896	6	107,169,822	A/T	0.147	intron	*LOC100422737*	0.030	0.0052	9.09E-09	3.5	3.93E-01	3.40E-01	1.01	0.96	1.06
rs4897160	6	126,223,944	A/G	0.470	intron	*NCOA7*	0.024	0.0031	2.04E-14	19.1	8.87E-02	9.70E-02	1.03	1.00	1.05
rs62435145	7	1,286,567	T/G	0.532	unknown	*UNCX*	0.040	0.0038	3.80E-26	11.7	2.39E-01	8.90E-01	1.06	1.03	1.10
rs12669187	7	30,915,478	A/G	0.135	intron	*INMT-MINDY4*	0.032	0.0057	1.25E-08	0	6.08E-01	6.10E-01	1.08	1.01	1.15
rs1051921	7	73,007,943	A/G	0.168	ncRNA, untranslated-3'	*MLXIPL*	-0.051	0.0042	2.86E-34	10.2	2.39E-01	6.50E-01	0.89	0.87	0.92
rs11551890	7	97,845,713	A/G	0.488	untranslated-3'	*TECPR1*	0.024	0.0040	2.05E-09	9.7	2.51E-01	4.50E-01	1.03	1.00	1.05
rs73728279	7	151,411,494	T/G	0.273	intron	*PRKAG2*	0.033	0.0043	1.15E-14	0.8	4.58E-01	4.50E-02	1.08	1.05	1.11
rs11781985	8	8,589,783	T/C	0.803	unknown	*CLDN23*	-0.025	0.0041	7.38E-10	12.3	1.97E-01	4.10E-01	0.99	0.96	1.02
rs34861762	8	23,748,420	T/C	0.366	unknown	*STC1*	0.034	0.0033	1.62E-25	24.5	3.31E-02	8.60E-02	1.02	1.00	1.05
rs7005606	8	32,401,501	T/G	0.597	intron	*NRG1*	-0.021	0.0032	9.56E-11	0	6.08E-01	9.90E-01	0.97	0.95	1.00
rs2941487	8	76,480,350	T/C	0.529	unknown	*HNF4G*	-0.043	0.0031	5.68E-44	13.6	1.68E-01	9.00E-01	0.95	0.92	0.97
rs62517932	8	77,031,593	A/G	0.079	unknown	*LINC01111*	0.041	0.0069	3.04E-09	8.4	2.78E-01	1.90E-01	1.01	0.96	1.06
rs12548367	8	95,929,202	T/C	0.679	unknown	*TP53INP1*	0.024	0.0033	4.39E-13	21	6.17E-02	2.60E-01	1.03	1.00	1.05

RSID	Chr.	position (b37)	A1/A2	EAF	Function	Gene Name	Effect (mg/dL)	SE	p-value	I2 (%)	P_{het}	$P_{anc-het}$	Gout OR	Gout OR 95% CI	Gout OR 95% CI
rs2436962	8	103,645,813	A/G	0.180	unknown	*KLF10*	0.026	0.0042	2.51E-10	0	9.43E-01	8.80E-01	1.03	0.99	1.07
rs3850445	9	16,703,381	A/G	0.384	intron	*BNC2*	0.021	0.0037	1.69E-08	0	9.94E-01	7.80E-01	1.03	0.99	1.06
rs7868781	9	33,141,320	A/G	0.609	intron	*B4GALT1*	0.030	0.0032	5.34E-20	20.3	6.94E-02	3.90E-01	1.05	1.02	1.08
rs3174352	9	95,219,248	A/G	0.464	intron, untranslated-3'	*CENPP*	-0.019	0.0032	3.41E-09	30.2	9.40E-03	7.00E-01	0.98	0.96	1.01
rs1800977	9	107,690,450	A/G	0.318	near-gene-5', untranslated-5'	*ABCA1*	-0.019	0.0034	7.95E-09	3.7	3.90E-01	7.10E-01	0.97	0.95	1.00
rs62585312	9	130,723,419	C/G	0.935	intron	*FAM102A*	0.053	0.0068	4.11E-15	0	8.41E-01	5.00E-01	1.07	1.01	1.14
rs10901057	9	134,245,698	C/G	0.258	unknown	*PRRC2B*	-0.026	0.0047	2.71E-08	0	9.12E-01	1.10E-01	0.92	0.88	0.96
rs74440730	10	16,920,892	A/C	0.894	intron	*CUBN*	-0.039	0.0059	5.50E-11	0	6.72E-01	6.70E-01	0.92	0.88	0.95
rs10994860	10	52,645,424	T/C	0.168	near-gene-5', untranslated-5'	*A1CF*	0.061	0.0045	7.44E-42	23.8	3.87E-02	2.80E-02	1.11	1.07	1.14
rs1649078	10	60,293,320	A/C	0.513	intron	*BICC1*	-0.045	0.0034	2.68E-39	25.1	4.37E-02	9.90E-02	0.94	0.92	0.97
rs1171617	10	61,467,182	T/G	0.770	intron	*SLC16A9*	0.081	0.0044	2.14E-77	0	9.04E-01	5.10E-01	1.15	1.12	1.19
rs9415676	10	65,010,626	A/G	0.650	intron	*JMJD1C*	-0.019	0.0033	4.53E-09	0	9.85E-01	9.30E-01	0.96	0.93	0.98
rs9420446	10	88,880,689	T/C	0.314	intron	*FAM35A*	-0.041	0.0044	1.37E-20	1.4	4.45E-01	8.20E-01	0.96	0.92	0.99
rs35198068	10	114,754,784	T/C	0.721	intron	*TCF7L2*	0.028	0.0038	2.57E-13	4.5	3.69E-01	1.90E-01	1.00	0.98	1.03
rs10886117	10	119,480,578	A/G	0.262	unknown	*EMX2*	0.028	0.0040	2.42E-12	14.8	1.49E-01	4.80E-01	1.03	1.00	1.07
rs4962699	10	126,477,209	A/G	0.250	intron	*EEF1AKMT2*	0.020	0.0036	1.65E-08	0	6.71E-01	5.40E-01	1.06	1.03	1.09
rs3741210	11	2,169,540	A/G	0.658	intron, ncRNA	*IGF2*	0.019	0.0034	1.17E-08	0	8.72E-01	1.50E-01	1.01	0.99	1.04
rs35229181	11	30,241,698	A/G	0.792	unknown	*FSHB*	0.022	0.0038	5.56E-09	10.7	2.26E-01	6.30E-01	1.02	0.99	1.05
rs963837	11	30,749,090	T/C	0.592	unknown	*DCDC1*	0.034	0.0033	4.75E-25	0	5.35E-01	5.00E-02	1.05	1.02	1.08
rs174594	11	61,619,829	A/C	0.606	intron	*FADS2*	-0.018	0.0033	4.88E-08	15.8	1.32E-01	9.50E-04	0.97	0.95	1.00
rs148838714	11	62,732,352	A/G	0.066	unknown	*SLC22A6*	-0.074	0.0091	2.14E-16	46.6	8.48E-05	1.40E-09	0.98	0.89	1.06
rs143825439	11	63,319,993	T/G	0.054	near-gene-3'	*HRASLS2*	-0.224	0.0121	1.23E-76	76.4	1.69E-17	4.80E-27	1.02	0.88	1.18
rs1006207	11	63,849,812	T/C	0.563	intron	*MACROD1*	-0.047	0.0035	3.54E-41	64.7	4.72E-14	3.20E-26	0.99	0.96	1.02
rs531763	11	64,352,063	A/G	0.565	unknown	*SLC22A12*	-0.116	0.0035	1.58E-246	83.7	2.30E-55	1.50E-67	0.86	0.84	0.89
rs34888828	11	64,968,104	A/G	0.110	intron	*CAPN1*	-0.057	0.0056	1.45E-23	75	6.91E-27	1.10E-49	1.02	0.97	1.07
rs4014195	11	65,506,822	C/G	0.688	unknown	*RNASEH2C*	-0.051	0.0034	6.87E-52	7.2	3.04E-01	2.90E-02	0.89	0.87	0.91
rs4073582	11	66,050,712	A/G	0.310	intron	*CNIH2*	-0.041	0.0037	5.41E-28	79.2	4.14E-37	1.00E-66	0.97	0.94	0.99
rs7110302	11	66,690,454	T/C	0.377	intron	*PC*	-0.018	0.0033	4.56E-08	5.3	3.51E-01	4.90E-01	0.97	0.94	0.99
rs11227805	11	67,246,757	T/C	0.180	unknown	*AIP*	-0.027	0.0046	8.72E-09	49.4	4.61E-06	4.00E-15	1.03	1.00	1.06
rs11217257	11	119,238,455	A/G	0.668	intron	*USP2*	-0.027	0.0035	2.81E-15	0	7.62E-01	4.60E-01	0.96	0.93	0.98
rs2058787	12	15,208,908	A/G	0.316	unknown	*LINC01489*	-0.020	0.0033	5.75E-09	0	5.05E-01	4.30E-01	0.97	0.94	0.99
rs4149056	12	21,331,549	T/C	0.847	missense	*SLCO1B1*	0.024	0.0043	2.22E-08	0	5.01E-01	2.00E-01	1.03	0.99	1.06
rs836968	12	50,267,335	T/C	0.371	intron	*FAIM2*	-0.020	0.0034	4.56E-09	0	8.07E-01	5.90E-01	0.99	0.96	1.02
rs11169926	12	52,248,787	A/C	0.344	intron	*LOC105369971*	0.030	0.0034	4.43E-18	0	4.90E-01	5.20E-01	1.05	1.02	1.08
rs73119306	12	57,826,982	A/G	0.786	near-gene-5'	*INHBC*	0.071	0.0042	7.45E-65	33.5	3.69E-03	6.80E-06	1.17	1.13	1.20
rs12368865	12	58,422,642	A/G	0.907	unknown	*LINC02403*	0.045	0.0067	1.31E-11	0	5.80E-01	7.70E-01	1.14	1.09	1.19
rs17550549	12	111,357,471	T/C	0.144	intron	*MYL2*	-0.035	0.0053	5.18E-11	33.9	3.29E-03	4.60E-09	1.05	1.00	1.11
rs10774625	12	111,910,219	A/G	0.467	intron	*ATXN2*	0.032	0.0038	1.29E-17	0	6.69E-01	8.40E-01	1.07	1.04	1.10
rs17696736	12	112,486,818	A/G	0.582	intron	*NAA25*	-0.028	0.0039	2.43E-13	0	5.21E-01	4.10E-01	0.95	0.92	0.97
rs11066390	12	113,163,766	A/G	0.268	unknown	*RPH3A*	0.025	0.0038	2.97E-11	6.6	3.19E-01	1.80E-01	1.05	1.02	1.08
rs1800574	12	121,416,864	T/C	0.033	missense, untranslated-5'	*HNF1A*	-0.085	0.0111	2.57E-14	20.8	7.24E-02	1.60E-01	0.92	0.85	0.99
rs148015593	12	122,523,668	T/G	0.495	unknown	*MLXIP*	0.029	0.0031	3.42E-21	14.8	1.49E-01	1.90E-01	1.06	1.04	1.09
rs74397112	12	133,094,111	T/C	0.148	intron	*FBRSL1*	0.043	0.0055	1.34E-14	14.3	1.73E-01	2.00E-03	1.07	1.03	1.11
rs9579574	13	31,025,634	A/G	0.293	unknown	*HMGB1*	-0.020	0.0036	2.03E-08	0	8.10E-01	1.50E-01	0.93	0.90	0.96
rs626277	13	72,347,696	A/C	0.499	intron	*DACH1*	0.025	0.0033	3.78E-14	1.6	4.39E-01	4.00E-01	1.02	1.00	1.05
rs8022225	14	55,767,069	A/G	0.556	intron	*FBXO34*	-0.018	0.0031	1.81E-08	2	4.29E-01	6.50E-01	0.97	0.95	1.00
rs861536	14	104,167,564	A/G	0.665	intron, near-gene-3', untranslated-3'	*KLC1*	0.021	0.0036	2.81E-09	13.1	1.79E-01	2.70E-01	1.05	1.02	1.08
rs11070231	15	40,021,576	A/C	0.570	intron	*FSIP1*	-0.028	0.0033	2.07E-17	0	5.54E-01	7.20E-01	0.97	0.95	1.00
rs2957740	15	72,307,691	A/G	0.412	intron	*MYO9A*	-0.024	0.0036	3.38E-11	9.7	2.52E-01	3.40E-01	0.98	0.96	1.01
rs8039418	15	73,441,432	T/C	0.471	intron	*NEO1*	-0.018	0.0032	2.39E-08	1	4.53E-01	1.40E-01	0.97	0.95	1.00

RSID	Chr.	position (b37)	A1/A2	EAF	Function	Gene Name	Effect (mg/dL)	SE	p-value	I2 (%)	P_{het}	$P_{anc-het}$	Gout OR	Gout OR 95% CI	Gout OR 95% CI
rs2472297	15	75,027,880	T/C	0.242	unknown	*CYP1A1*	-0.028	0.0048	8.49E-09	0	7.87E-01	9.70E-01	0.96	0.93	0.99
rs73436803	15	75,619,201	T/C	0.244	unknown	*COMMD4*	-0.027	0.0039	2.58E-12	0	4.77E-01	1.70E-02	0.96	0.93	0.98
rs11072567	15	76,298,744	A/G	0.500	intron	*NRG4*	-0.042	0.0032	2.52E-39	25	3.08E-02	9.60E-02	0.93	0.90	0.95
rs494268	15	76,815,713	T/C	0.908	intron	*SCAPER*	0.032	0.0054	3.96E-09	0	9.87E-01	3.00E-01	1.09	1.04	1.14
rs8024386	15	90,670,526	A/C	0.774	unknown	*IDH2*	-0.022	0.0039	2.12E-08	12.3	1.97E-01	3.30E-02	0.92	0.89	0.95
rs12908437	15	99,287,375	T/C	0.433	intron	*IGF1R*	0.045	0.0031	3.77E-46	0	6.90E-01	1.90E-01	1.05	1.03	1.08
rs77924615	16	20,392,332	A/G	0.203	intron	*PDILT*	-0.027	0.0040	1.27E-11	23	4.60E-02	9.60E-01	0.98	0.95	1.01
rs2219647	16	51,733,405	A/G	0.266	unknown	*LINC01571*	0.021	0.0035	5.38E-09	12.8	1.83E-01	9.40E-01	1.04	1.01	1.07
rs62033406	16	53,824,226	A/G	0.553	intron	*FTO*	-0.019	0.0032	3.23E-09	9.5	2.52E-01	1.30E-02	0.96	0.93	0.98
rs56230350	16	68,166,971	A/C	0.873	intron	*NFATC3*	0.026	0.0046	3.08E-08	0	4.99E-01	2.30E-01	1.02	0.98	1.06
rs62052820	16	69,575,238	A/G	0.194	unknown	*MIR1538*	0.043	0.0040	1.17E-26	2	4.29E-01	7.10E-01	1.08	1.05	1.12
rs4788815	16	71,634,811	A/T	0.343	unknown	*TAT*	-0.024	0.0033	2.93E-13	0	6.73E-01	6.80E-01	0.98	0.95	1.00
rs9302635	16	72,144,174	T/C	0.763	intron	*DHX38*	-0.028	0.0046	2.24E-09	0	8.88E-01	5.40E-01	0.97	0.94	1.00
rs57652769	16	79,753,976	T/C	0.295	unknown	*MAFTRR*	-0.037	0.0034	2.32E-27	0	5.74E-01	9.40E-01	0.96	0.93	0.98
rs11644696	16	81,572,093	A/G	0.408	intron	*CMIP*	0.019	0.0033	6.42E-09	0	6.97E-01	5.10E-01	1.01	0.99	1.04
rs7212936	17	1,646,651	A/C	0.452	intron, near-gene-5'	*SERPINF2*	-0.018	0.0032	1.75E-08	8.2	2.87E-01	2.80E-01	0.97	0.95	1.00
rs2453580	17	19,438,321	T/C	0.639	intron	*SLC47A1*	0.022	0.0035	3.49E-10	0	4.84E-01	2.50E-02	1.04	1.01	1.06
rs7209801	17	42,323,376	A/G	0.403	unknown	*SLC4A1*	0.019	0.0033	5.19E-09	0	9.24E-01	2.00E-01	1.03	1.00	1.06
rs3794748	17	53,365,172	A/G	0.352	intron	*HLF*	0.033	0.0034	8.68E-23	15.3	1.42E-01	7.50E-02	1.05	1.03	1.08
rs2645477	17	57,845,624	A/C	0.521	intron	*VMP1*	0.018	0.0031	7.66E-09	12.6	1.89E-01	7.20E-02	1.02	1.00	1.05
rs9895661	17	59,456,589	T/C	0.674	intron	*BCAS3*	0.053	0.0038	7.63E-45	28.5	1.54E-02	2.80E-02	1.09	1.06	1.13
rs164011	17	74,273,165	A/G	0.411	intron	*QRICH2*	-0.021	0.0033	1.05E-10	0	7.91E-01	9.10E-01	0.96	0.93	0.98
rs10438961	18	42,779,107	T/G	0.730	unknown	*SLC14A2*	-0.021	0.0037	2.01E-08	0	9.80E-01	9.40E-02	0.97	0.94	0.99
rs11663816	18	57,876,227	T/C	0.754	unknown	*MC4R*	-0.026	0.0036	1.25E-12	6	3.34E-01	7.30E-02	0.93	0.91	0.96
rs7259484	19	1,813,207	A/G	0.653	near-gene-5'	*ATP8B3*	0.021	0.0036	3.60E-09	0	8.55E-01	1.50E-01	1.01	0.98	1.05
rs4807003	19	4,957,133	A/G	0.316	intron	*UHRF1*	-0.028	0.0041	8.85E-12	6.5	3.23E-01	5.00E-01	0.97	0.95	1.00
rs10405423	19	7,211,311	A/C	0.700	intron	*INSR*	0.033	0.0037	1.11E-18	0	6.35E-01	1.20E-02	1.06	1.03	1.09
rs56338130	19	18,235,871	T/C	0.224	intron	*MAST3*	0.023	0.0038	4.15E-09	13.1	1.79E-01	6.70E-01	1.03	1.00	1.06
rs10418164	19	33,411,139	T/G	0.455	intron	*CEP89*	0.025	0.0032	3.14E-15	3.6	3.91E-01	7.50E-01	1.04	1.02	1.07
rs62128132	19	50,217,955	T/C	0.967	unknown	*CPT1C*	-0.115	0.0144	1.99E-15	72.7	7.42E-16	3.00E-07	0.78	0.72	0.84
rs12625256	20	10,638,386	A/T	0.594	intron	*JAG1*	0.019	0.0032	1.93E-09	0	7.01E-01	3.90E-01	1.03	1.00	1.05
rs6138584	20	25,463,148	A/T	0.236	intron	*NINL*	0.023	0.0038	2.50E-09	0	9.91E-01	9.40E-01	1.02	0.99	1.06
rs6119524	20	33,373,813	T/C	0.343	intron	*NCOA6*	0.020	0.0033	3.50E-09	19.9	7.43E-02	4.00E-01	1.00	0.97	1.03
rs73611258	20	39,742,284	A/G	0.648	intron	*TOP1*	0.024	0.0039	4.03E-10	7.3	3.01E-01	3.70E-01	1.01	0.97	1.04
rs6031598	20	43,056,149	T/G	0.535	intron	*HNF4A*	-0.022	0.0032	1.73E-12	9.4	2.58E-01	9.60E-01	0.95	0.92	0.97
rs1407040	20	57,472,174	T/C	0.685	intron	*GNAS*	-0.025	0.0033	5.70E-14	23	4.37E-02	4.40E-02	0.98	0.95	1.00
rs2834319	21	35,357,025	T/C	0.851	unknown	*LOC101928126*	-0.026	0.0045	8.90E-09	0	8.37E-01	7.60E-01	0.98	0.94	1.01
rs219781	21	37,832,621	T/G	0.247	near-gene-3'	*CLDN14*	-0.027	0.0042	2.22E-10	0	8.64E-01	5.30E-01	0.95	0.93	0.98
rs12484809	22	44,325,631	T/C	0.284	intron	*PNPLA3*	-0.035	0.0037	1.73E-20	0	5.84E-01	9.20E-01	0.91	0.88	0.94

Supplementary Table 5 – MR-MEGA SNPs not identified as METAL index SNPs with the filter on number of cohorts > 37 relaxed.

** denotes a P-value significant at P < 5E-8 and the * symbol significant at P < 0.05/22.

rsID	Chr.	Position	Gene	Function	EA/NEA	EAF	N	N cohorts	N SNPs	N Eth	Effect	SE	P	$P_{anc-het}$	$P_{res-het}$	METAL P
rs508926	1	28,578,825	*RP5-1092A3.5*	Downstream	G/A	0.280	447,631	73	1	5	0.034	0.034	4.69E-08	2.03E-05*	1.91E-01	2.68E-04*
rs186147970	1	120,484,718	*NOTCH2*	Intron	C/T	0.006	66,139	8	1	1	1.586	2.645	4.13E-14	9.22E-01	6.94E-01	1.46E-16**
rs188991775	1	147,336,171	*RP11-314N2.2*	Intergenic	T/C	0.995	66,369	6	2	1	-1.282	4.208	9.16E-12	1.12E-03*	4.79E-01	8.93E-11**
rs3774046	3	170,737,003	*SLC2A2*	Intron	G/A	0.160	457,616	74	18	5	0.050	0.048	1.84E-09	4.40E-04*	1.61E-01	8.72E-07*
rs17325213	4	11,955,802	*TAPT1-AS1*	Intergenic	T/C	0.056	239,512	48	4	4	1.481	0.258	3.59E-09	5.51E-08*	3.78E-01	3.49E-03
rs550638591	4	36,873,382	*RP11-500G9.1*	Intergenic	C/T	0.003	24,997	6	2	1	28.806	5.659	1.45E-10	4.88E-11**	9.79E-01	5.35E-01
rs1346144	4	79,625,361	*RP11-576N17.4*	Intergenic	A/G	0.631	422,278	72	1	5	-0.186	0.052	2.24E-08	7.53E-03	8.94E-01	2.65E-07*
rs73728140	6	24,507,003	*ALDH5A1*	Intron	G/A	0.046	445,653	69	1	5	0.064	0.056	1.03E-08	2.88E-09**	5.14E-01	3.68E-01
rs570169004	9	70,930,514	*RP11-561O23.5*	Intron	G/A	0.010	60,114	6	1	1	-12.620	81.275	7.66E-10	3.83E-10**	2.17E-03*	2.01E-01
rs697238	10	80,947,668	*ZMIZ1*	Intron	G/T	0.614	456,290	73	5	5	0.026	0.034	2.46E-09	1.37E-03*	2.91E-02	3.11E-07*
rs60808706	11	2,857,233	*KCNQ1*	Intron	A/G	0.160	446,270	72	1	5	-0.013	0.055	1.67E-08	5.86E-02	7.53E-01	8.56E-08*
rs334	11	5,248,232	*Hb*	Missense	A/T	0.060	34,279	8	3	2	0.057	0.726	3.01E-13	4.65E-01	4.24E-01	5.28E-15**
rs11601310	11	48,085,189	*PTPRJ*	Intron	A/G	0.236	443,108	69	1	5	-0.011	0.039	8.60E-09	3.40E-05*	4.12E-01	5.04E-06*
rs118077950	11	67,902,411	*CTD-2655K5.1*	Downstream	A/G	0.105	143,037	9	10	4	-0.697	0.661	5.89E-12	1.37E-01	3.73E-02	3.14E-11**
rs186255333	11	72,766,638	*FCHSD2*	Intron	C/T	0.003	23,901	6	1	1	54.652	50.188	2.27E-08	2.42E-06*	2.77E-01	4.96E-03
rs75198898	11	116,649,806	*ZPR1*	Intron	A/G	0.058	166,398	9	2	3	1.434	0.245	2.89E-09	1.99E-06*	6.75E-01	8.51E-05*
rs144074240	12	67,657,757	*GGTA2P*	Downstream	T/C	0.061	330,887	70	3	4	0.121	0.127	1.60E-08	1.33E-03*	9.68E-01	3.48E-07*
rs145995319	14	98,100,885	*RP11-76E12.1*	Intron	A/G	0.995	126,986	32	1	3	-7.805	2.311	3.18E-08	2.95E-08**	3.41E-05*	9.58E-02
rs73051934	19	48,582,916	*PLA2G4C*	Intron	G/A	0.015	227,758	33	1	2	-5.485	0.833	2.76E-11	1.28E-11**	2.13E-01	1.70E-01
rs200771360	19	49,129,618	*SPHK2*	Splice region	T/G	0.009	69,248	7	5	1	-13.192	3.692	6.64E-23	6.97E-05*	2.38E-01	2.44E-17**
rs563655432	19	49,634,571	*PPFIA3*	Intron	T/G	0.009	63,227	6	8	1	-17.704	0.398	1.09E-44	1.11E-03	9.96E-01	1.93E-44
rs562175867	19	51,044,313	*LRRC4B*	Intron	C/T	0.012	64,313	6	6	1	-21.339	6.681	1.77E-28	1.96E-05	1.75E-01	1.78E-21

Supplementary Table 6 – Spearman correlations between DEPICT exemplar gene sets (r > 0.2)

Source Node	Target Node	Correlation Coefficient
HEBP1 PPI	HEBP1 PPI	1
ABCC9 PPI	ABCC9 PPI	1
FLT1 PPI	FLT1 PPI	1
UBE2I PPI	UBE2I PPI	1
CRTC1 PPI	CRTC1 PPI	1
SERPINE1 PPI	SERPINE1 PPI	1
Response to Hypoxia	SERPINE1 PPI	0.265213
YWHAE PPI	YWHAE PPI	1
NR3C1 PPI	NR3C1 PPI	1
FHL2 PPI	FHL2 PPI	1
PLG PPI	PLG PPI	1
NCOA3 PPI	NCOA3 PPI	1
ACVR1B PPI	ACVR1B PPI	1
Transforming Growth Factor Beta Receptor Cytoplasmic Mediator Activity	ACVR1B PPI	0.20153
PDGFC PPI	PDGFC PPI	1
ITGA5 PPI	ITGA5 PPI	1
ENSG00000215320 PPI	ENSG00000215320 PPI	1
Abnormal Long Bone Epiphyseal Plate Morphology	ENSG00000215320 PPI	0.242783
Abnormal Bone Ossification	ENSG00000215320 PPI	0.339356
SERPINE1 PPI	Response to Hypoxia	0.265213
Response to Hypoxia	Response to Hypoxia	1
Glucose Catabolic Process	Response to Hypoxia	0.240306
Kidney Development	Kidney Development	1
Regulation of Organ Morphogenesis	Kidney Development	0.361279
Enlarged Kidney	Kidney Development	0.227648
Thoracic Vertebral Transformation	Kidney Development	0.220185
Increased Glomerular Capsule Space	Kidney Development	0.292251
Liver Development	Liver Development	1
Steroid Hormone Receptor Activity	Liver Development	0.218233
Increased Glomerular Capsule Space	Liver Development	0.207807
Liver Development	Steroid Hormone Receptor Activity	0.218233
Steroid Hormone Receptor Activity	Steroid Hormone Receptor Activity	1
Decreased Circulating Cholesterol Level	Steroid Hormone Receptor Activity	0.217517
ACVR1B PPI	Transforming Growth Factor Beta Receptor Cytoplasmic Mediator Activity	0.20153
Transforming Growth Factor Beta Receptor Cytoplasmic Mediator Activity	Transforming Growth Factor Beta Receptor Cytoplasmic Mediator Activity	1
Microsome	Microsome	1
Electron Carrier Activity	Microsome	0.223806
Ppara Activates Gene Expression	Microsome	0.608128
Response to Hypoxia	Glucose Catabolic Process	0.240306
Glucose Catabolic Process	Glucose Catabolic Process	1
Hexose Metabolic Process	Glucose Catabolic Process	0.376677
Regulation of Carbohydrate Metabolic Process	Regulation of Carbohydrate Metabolic Process	1
Microsome	Electron Carrier Activity	0.223806
Electron Carrier Activity	Electron Carrier Activity	1
Glucose Catabolic Process	Hexose Metabolic Process	0.376677
Hexose Metabolic Process	Hexose Metabolic Process	1
Cell-Substrate Junction	Cell-Substrate Junction	1
Respiratory Tube Development	Respiratory Tube Development	1
Tube Development	Respiratory Tube Development	0.301647
Mesangial Cell Hyperplasia	Respiratory Tube Development	0.212057
Response to Nutrient Levels	Response to Nutrient Levels	1
Decreased Circulating Cholesterol Level	Response to Nutrient Levels	0.20028
Nuclear Hormone Receptor Binding	Nuclear Hormone Receptor Binding	1
Respiratory Tube Development	Tube Development	0.301647
Tube Development	Tube Development	1
Wnt-Activated Receptor Activity	Wnt-Activated Receptor Activity	1
Abnormal Long Bone Epiphyseal Plate Morphology	Wnt-Activated Receptor Activity	0.463033
Response to Peptide Hormone Stimulus	Response to Peptide Hormone Stimulus	1
Abnormal Glucose Homeostasis	Response to Peptide Hormone Stimulus	0.426309
Small Molecule Catabolic Process	Small Molecule Catabolic Process	1
Apical Part of Cell	Apical Part of Cell	1
Hydronephrosis	Apical Part of Cell	0.278573
Abnormal Kidney Physiology	Apical Part of Cell	0.508863
Decreased Urine Osmolality	Apical Part of Cell	0.485054
Enlarged Kidney	Apical Part of Cell	0.297566
Increased Urine Calcium Level	Apical Part of Cell	0.74445
Increased Circulating Potassium Level	Apical Part of Cell	0.245951

Source Node	Target Node	Correlation Coefficient
Increased Urine Sodium Level	Apical Part of Cell	0.465027
Kidney Cortex Atrophy	Apical Part of Cell	0.361807
Slc-Mediated Transmembrane Transport	Apical Part of Cell	0.759477
Negative Regulation of Growth	Negative Regulation of Growth	1
Embryonic Morphogenesis	Embryonic Morphogenesis	1
Thoracic Vertebral Transformation	Embryonic Morphogenesis	0.892143
Kidney Development	Regulation of Organ Morphogenesis	0.361279
Regulation of Organ Morphogenesis	Regulation of Organ Morphogenesis	1
Thoracic Vertebral Transformation	Regulation of Organ Morphogenesis	0.231187
Apical Part of Cell	Hydronephrosis	0.278573
Hydronephrosis	Hydronephrosis	1
Abnormal Kidney Physiology	Hydronephrosis	0.51184
Increased Circulating Aldosterone Level	Hydronephrosis	0.337437
Decreased Urine Osmolality	Hydronephrosis	0.464381
Enlarged Kidney	Hydronephrosis	0.330851
Increased Urine Calcium Level	Hydronephrosis	0.400911
Increased Circulating Potassium Level	Hydronephrosis	0.469428
Increased Urine Sodium Level	Hydronephrosis	0.616328
Kidney Cortex Atrophy	Hydronephrosis	0.388661
Mesangial Cell Hyperplasia	Hydronephrosis	0.288615
Decreased Embryo Size	Decreased Embryo Size	1
Response to Peptide Hormone Stimulus	Abnormal Glucose Homeostasis	0.426309
Abnormal Glucose Homeostasis	Abnormal Glucose Homeostasis	1
Apical Part of Cell	Abnormal Kidney Physiology	0.508863
Hydronephrosis	Abnormal Kidney Physiology	0.51184
Abnormal Kidney Physiology	Abnormal Kidney Physiology	1
Increased Circulating Aldosterone Level	Abnormal Kidney Physiology	0.213069
Decreased Urine Osmolality	Abnormal Kidney Physiology	0.758218
Enlarged Kidney	Abnormal Kidney Physiology	0.457083
Increased Urine Calcium Level	Abnormal Kidney Physiology	0.709209
Increased Circulating Potassium Level	Abnormal Kidney Physiology	0.43471
Increased Urine Sodium Level	Abnormal Kidney Physiology	0.797942
Kidney Cortex Atrophy	Abnormal Kidney Physiology	0.414245
Slc-Mediated Transmembrane Transport	Abnormal Kidney Physiology	0.408202
Hydronephrosis	Increased Circulating Aldosterone Level	0.337437
Abnormal Kidney Physiology	Increased Circulating Aldosterone Level	0.213069
Increased Circulating Aldosterone Level	Increased Circulating Aldosterone Level	1
Increased Circulating Potassium Level	Increased Circulating Aldosterone Level	0.55589
Increased Urine Sodium Level	Increased Circulating Aldosterone Level	0.395197
Apical Part of Cell	Decreased Urine Osmolality	0.485054
Hydronephrosis	Decreased Urine Osmolality	0.464381
Abnormal Kidney Physiology	Decreased Urine Osmolality	0.758218
Decreased Urine Osmolality	Decreased Urine Osmolality	1
Enlarged Kidney	Decreased Urine Osmolality	0.571831
Increased Urine Calcium Level	Decreased Urine Osmolality	0.818558
Increased Circulating Potassium Level	Decreased Urine Osmolality	0.330216
Increased Urine Sodium Level	Decreased Urine Osmolality	0.70046
Kidney Cortex Atrophy	Decreased Urine Osmolality	0.323056
Mesangial Cell Hyperplasia	Decreased Urine Osmolality	0.208061
Slc-Mediated Transmembrane Transport	Decreased Urine Osmolality	0.456386
ENSG00000215320 PPI	Abnormal Long Bone Epiphyseal Plate Morphology	0.242783
Wnt-Activated Receptor Activity	Abnormal Long Bone Epiphyseal Plate Morphology	0.463033
Abnormal Long Bone Epiphyseal Plate Morphology	Abnormal Long Bone Epiphyseal Plate Morphology	1
Abnormal Bone Ossification	Abnormal Long Bone Epiphyseal Plate Morphology	0.662422
Kidney Development	Enlarged Kidney	0.227648
Apical Part of Cell	Enlarged Kidney	0.297566
Hydronephrosis	Enlarged Kidney	0.330851
Abnormal Kidney Physiology	Enlarged Kidney	0.457083
Decreased Urine Osmolality	Enlarged Kidney	0.571831
Enlarged Kidney	Enlarged Kidney	1
Increased Urine Calcium Level	Enlarged Kidney	0.555349
Increased Urine Sodium Level	Enlarged Kidney	0.521318
Kidney Cortex Atrophy	Enlarged Kidney	0.309675
Mesangial Cell Hyperplasia	Enlarged Kidney	0.23306
Slc-Mediated Transmembrane Transport	Enlarged Kidney	0.393148
Kidney Development	Thoracic Vertebral Transformation	0.220185
Embryonic Morphogenesis	Thoracic Vertebral Transformation	0.892143
Regulation of Organ Morphogenesis	Thoracic Vertebral Transformation	0.231187
Thoracic Vertebral Transformation	Thoracic Vertebral Transformation	1
Steroid Hormone Receptor Activity	Decreased Circulating Cholesterol Level	0.217517
Response to Nutrient Levels	Decreased Circulating Cholesterol Level	0.20028
Decreased Circulating Cholesterol Level	Decreased Circulating Cholesterol Level	1
Apical Part of Cell	Increased Urine Calcium Level	0.74445

Source Node	Target Node	Correlation Coefficient
Hydronephrosis	Increased Urine Calcium Level	0.400911
Abnormal Kidney Physiology	Increased Urine Calcium Level	0.709209
Decreased Urine Osmolality	Increased Urine Calcium Level	0.818558
Enlarged Kidney	Increased Urine Calcium Level	0.555349
Increased Urine Calcium Level	Increased Urine Calcium Level	1
Increased Circulating Potassium Level	Increased Urine Calcium Level	0.220863
Increased Urine Sodium Level	Increased Urine Calcium Level	0.661588
Kidney Cortex Atrophy	Increased Urine Calcium Level	0.414838
Slc-Mediated Transmembrane Transport	Increased Urine Calcium Level	0.682957
Apical Part of Cell	Increased Circulating Potassium Level	0.245951
Hydronephrosis	Increased Circulating Potassium Level	0.469428
Abnormal Kidney Physiology	Increased Circulating Potassium Level	0.43471
Increased Circulating Aldosterone Level	Increased Circulating Potassium Level	0.55589
Decreased Urine Osmolality	Increased Circulating Potassium Level	0.330216
Increased Urine Calcium Level	Increased Circulating Potassium Level	0.220863
Increased Circulating Potassium Level	Increased Circulating Potassium Level	1
Increased Urine Sodium Level	Increased Circulating Potassium Level	0.504154
Apical Part of Cell	Increased Urine Sodium Level	0.465027
Hydronephrosis	Increased Urine Sodium Level	0.616328
Abnormal Kidney Physiology	Increased Urine Sodium Level	0.797942
Increased Circulating Aldosterone Level	Increased Urine Sodium Level	0.395197
Decreased Urine Osmolality	Increased Urine Sodium Level	0.70046
Enlarged Kidney	Increased Urine Sodium Level	0.521318
Increased Urine Calcium Level	Increased Urine Sodium Level	0.661588
Increased Circulating Potassium Level	Increased Urine Sodium Level	0.504154
Increased Urine Sodium Level	Increased Urine Sodium Level	1
Kidney Cortex Atrophy	Increased Urine Sodium Level	0.281527
Slc-Mediated Transmembrane Transport	Increased Urine Sodium Level	0.298183
ENSG00000215320 PPI	Abnormal Bone Ossification	0.339356
Abnormal Long Bone Epiphyseal Plate Morphology	Abnormal Bone Ossification	0.662422
Abnormal Bone Ossification	Abnormal Bone Ossification	1
Partial Prenatal Lethality	Partial Prenatal Lethality	1
Apical Part of Cell	Kidney Cortex Atrophy	0.361807
Hydronephrosis	Kidney Cortex Atrophy	0.388661
Abnormal Kidney Physiology	Kidney Cortex Atrophy	0.414245
Decreased Urine Osmolality	Kidney Cortex Atrophy	0.323056
Enlarged Kidney	Kidney Cortex Atrophy	0.309675
Increased Urine Calcium Level	Kidney Cortex Atrophy	0.414838
Increased Urine Sodium Level	Kidney Cortex Atrophy	0.281527
Kidney Cortex Atrophy	Kidney Cortex Atrophy	1
Mesangial Cell Hyperplasia	Kidney Cortex Atrophy	0.23662
Slc-Mediated Transmembrane Transport	Kidney Cortex Atrophy	0.383702
Respiratory Tube Development	Mesangial Cell Hyperplasia	0.212057
Hydronephrosis	Mesangial Cell Hyperplasia	0.288615
Decreased Urine Osmolality	Mesangial Cell Hyperplasia	0.208061
Enlarged Kidney	Mesangial Cell Hyperplasia	0.23306
Kidney Cortex Atrophy	Mesangial Cell Hyperplasia	0.23662
Mesangial Cell Hyperplasia	Mesangial Cell Hyperplasia	1
Increased Glomerular Capsule Space	Mesangial Cell Hyperplasia	0.278367
Kidney Development	Increased Glomerular Capsule Space	0.292251
Liver Development	Increased Glomerular Capsule Space	0.207807
Mesangial Cell Hyperplasia	Increased Glomerular Capsule Space	0.278367
Increased Glomerular Capsule Space	Increased Glomerular Capsule Space	1
Microsome	Ppara Activates Gene Expression	0.608128
Ppara Activates Gene Expression	Ppara Activates Gene Expression	1
Apical Part of Cell	Slc-Mediated Transmembrane Transport	0.759477
Abnormal Kidney Physiology	Slc-Mediated Transmembrane Transport	0.408202
Decreased Urine Osmolality	Slc-Mediated Transmembrane Transport	0.456386
Enlarged Kidney	Slc-Mediated Transmembrane Transport	0.393148
Increased Urine Calcium Level	Slc-Mediated Transmembrane Transport	0.682957
Increased Urine Sodium Level	Slc-Mediated Transmembrane Transport	0.298183
Kidney Cortex Atrophy	Slc-Mediated Transmembrane Transport	0.383702
Slc-Mediated Transmembrane Transport	Slc-Mediated Transmembrane Transport	1

Supplementary Table 7 – All sex-separate genetic correlations significant in males or females.

Highlighted cells are significant (Q < 0.05) or suggestive in the case of $P_{sex\text{-}diff}$ (P< 0.05).

Trait	PMID	Category	r_g Male	r_g Female	SE Male	SE Female	P-value Male	P-value Female	Q-value Male	Q-value Female	Normalised Difference Score	$P_{sex\text{-}diff}$	$Q_{sex\text{-}diff}$
Parents age at death	27015805	aging	-0.225	-0.151	0.076	0.069	3.00E-03	2.88E-02	1.19E-02	7.81E-02	0.713752	0.475	0.669
Fathers age at death	27015805	aging	-0.271	-0.278	0.066	0.073	3.63E-05	1.00E-04	2.43E-04	5.52E-04	-0.07319	0.942	0.973
Child birth weight	23202124	anthropometric	-0.048	-0.125	0.054	0.051	3.68E-01	1.36E-02	5.74E-01	4.16E-02	-1.03506	0.301	0.498
Body mass index	20935630	anthropometric	0.196	0.330	0.041	0.049	1.51E-06	2.11E-11	1.47E-05	8.94E-10	2.102316	0.036	0.258
Body fat	26833246	anthropometric	0.235	0.323	0.052	0.051	6.18E-06	1.81E-10	4.62E-05	5.74E-09	1.210316	0.226	0.433
Childhood obesity	22484627	anthropometric	0.102	0.235	0.045	0.050	2.40E-02	2.99E-06	6.70E-02	2.71E-05	1.972799	0.049	0.286
Extreme bmi	23563607	anthropometric	0.251	0.313	0.062	0.064	5.48E-05	1.16E-06	3.24E-04	1.22E-05	0.698065	0.485	0.669
Obesity class 1	23563607	anthropometric	0.192	0.318	0.042	0.062	4.13E-06	2.49E-07	3.39E-05	3.35E-06	1.687119	0.092	0.404
Obesity class 2	23563607	anthropometric	0.225	0.345	0.055	0.067	4.34E-05	2.51E-07	2.81E-04	3.35E-06	1.388964	0.165	0.428
Obesity class 3	23563607	anthropometric	0.181	0.310	0.078	0.083	2.09E-02	2.00E-04	6.10E-02	1.04E-03	1.129012	0.259	0.448
rOverweight	23563607	anthropometric	0.193	0.311	0.044	0.048	1.39E-05	7.17E-11	9.79E-05	2.60E-09	1.821347	0.069	0.363
Hip circumference	25673412	anthropometric	0.215	0.313	0.042	0.054	2.78E-07	8.45E-09	3.53E-06	1.79E-07	1.427208	0.154	0.428
Waist circumference	25673412	anthropometric	0.243	0.382	0.048	0.071	2.98E-07	8.23E-08	3.61E-06	1.31E-06	1.62168	0.105	0.409
Waist-to-hip ratio	25673412	anthropometric	0.186	0.329	0.040	0.052	3.10E-06	2.40E-10	2.71E-05	6.78E-09	2.192931	0.028	0.258
Birth weight	27680694	anthropometric	-0.029	-0.090	0.034	0.030	3.91E-01	2.50E-03	5.94E-01	1.04E-02	-1.35373	0.176	0.428
Ulcerative colitis	26192919	autoimmune	-0.139	-0.068	0.044	0.040	1.70E-03	8.70E-02	7.32E-03	1.81E-01	1.203485	0.229	0.433
Femoral Neck bone mineral density	26367794	bone	0.042	0.120	0.046	0.036	3.57E-01	8.00E-04	5.66E-01	3.69E-03	1.332378	0.183	0.428
Lumbar Spine bone mineral density	26367794	bone	0.020	0.095	0.045	0.036	6.47E-01	8.50E-03	7.94E-01	2.88E-02	1.303631	0.192	0.428
Coronary artery disease	26343387	cardiometabolic	0.106	0.220	0.032	0.045	9.00E-04	9.14E-07	4.01E-03	1.01E-05	2.067604	0.039	0.258
Intelligence	28530673	cognitive	-0.042	-0.081	0.035	0.030	2.27E-01	6.60E-03	4.05E-01	2.36E-02	-0.8473	0.397	0.601
Years of schooling 2016	27225129	education	-0.062	-0.146	0.024	0.031	9.90E-03	3.29E-06	3.26E-02	2.78E-05	-2.13298	0.033	0.258
College completion	23722424	education	-0.079	-0.129	0.031	0.032	9.60E-03	4.98E-05	3.21E-02	3.08E-04	-1.12121	0.262	0.448
Years of schooling (proxy cognitive performance)	25201988	education	-0.089	-0.152	0.033	0.035	7.40E-03	1.51E-05	2.57E-02	1.04E-04	-1.31006	0.190	0.428
Years of schooling 2013	23722424	education	-0.069	-0.156	0.033	0.039	3.82E-02	5.31E-05	9.42E-02	3.21E-04	-1.71967	0.085	0.404
Type 2 Diabetes	22885922	glycemic	0.141	0.329	0.053	0.072	7.30E-03	5.24E-06	2.57E-02	4.16E-05	2.10235	0.036	0.258
Fasting glucose main effect	22581228	glycemic	0.146	0.208	0.057	0.062	1.00E-02	9.00E-04	3.26E-02	4.01E-03	0.725926	0.468	0.669
Fasting insulin main effect	22581228	glycemic	0.282	0.397	0.057	0.082	8.43E-07	1.27E-06	9.73E-06	1.29E-05	1.15224	0.249	0.448
HbA1C	20858683	glycemic	-0.005	0.168	0.057	0.062	9.24E-01	6.60E-03	9.54E-01	2.36E-02	2.065205	0.039	0.258
HOMA-B	20081858	glycemic	0.186	0.302	0.055	0.057	8.00E-04	1.32E-07	3.69E-03	1.97E-06	1.455417	0.146	0.428
HOMA-IR	20081858	glycemic	0.319	0.474	0.067	0.085	2.14E-06	2.58E-08	2.01E-05	5.04E-07	1.422285	0.155	0.428
Leptin_adjBMI	26833098	hormone	0.138	0.137	0.070	0.056	4.85E-02	1.46E-02	1.15E-01	4.36E-02	-0.01342	0.989	0.989
Leptin_not_adjBMI	26833098	hormone	0.273	0.367	0.076	0.067	3.00E-04	4.37E-08	1.49E-03	7.40E-07	0.93145	0.352	0.565

Trait	PMID	Category	r_g Male	r_g Female	SE Male	SE Female	P-value Male	P-value Female	Q-value Male	Q-value Female	Normalised Difference Score	$P_{sex\text{-}diff}$	$Q_{sex\text{-}diff}$
Chronic Kidney Disease	26831199	kidney	0.234	0.318	0.093	0.086	1.18E-02	2.00E-04	3.75E-02	1.04E-03	0.666792	0.505	0.669
Serum creatinine (non-diabetes)	26831199	kidney	-0.187	-0.208	0.047	0.046	7.29E-05	5.47E-06	4.11E-04	4.21E-05	-0.32386	0.746	0.920
Serum creatinine	26831199	kidney	-0.185	-0.200	0.049	0.046	2.00E-04	1.14E-05	1.04E-03	8.27E-05	-0.22931	0.819	0.933
Serum cystatin c	26831199	kidney	-0.287	-0.321	0.101	0.106	4.60E-03	2.40E-03	1.69E-02	1.02E-02	-0.22598	0.821	0.933
HDL cholesterol	20686565	lipids	-0.320	-0.324	0.047	0.054	5.99E-12	1.46E-09	3.34E-10	3.71E-08	-0.05919	0.953	0.973
Triglycerides	20686565	lipids	0.341	0.288	0.057	0.053	2.73E-09	4.32E-08	6.30E-08	7.40E-07	-0.68648	0.492	0.669
Forced expiratory volume in 1 second (FEV1)	28166213	lung_function	-0.080	-0.073	0.032	0.032	1.20E-02	2.13E-02	3.76E-02	6.15E-02	0.171754	0.864	0.934
Forced Vital capacity(FVC)	28166213	lung_function	-0.093	-0.102	0.031	0.030	2.90E-03	8.00E-04	1.17E-02	3.69E-03	-0.20496	0.838	0.933
Forced expiratory volume in 1 second (FEV1)	21946350	lung_function	-0.127	-0.112	0.050	0.046	1.14E-02	1.44E-02	3.67E-02	4.35E-02	0.208986	0.834	0.933
Anorexia Nervosa	24514567	psychiatric	-0.036	-0.099	0.033	0.034	2.74E-01	3.80E-03	4.71E-01	1.44E-02	-1.32561	0.185	0.428
Schizophrenia	25056061	psychiatric	-0.052	-0.087	0.027	0.028	5.19E-02	1.70E-03	1.22E-01	7.32E-03	-0.90127	0.367	0.573
Age at Menarche	25231870	reproductive	-0.097	-0.124	0.032	0.030	2.90E-03	4.42E-05	1.17E-02	2.81E-04	-0.62347	0.533	0.689
Age at Menopause	26414677	reproductive	0.015	-0.115	0.038	0.041	7.00E-01	4.50E-03	8.04E-01	1.68E-02	-2.3311	0.020	0.258
Age of first birth	27798627	reproductive	-0.107	-0.180	0.036	0.045	3.20E-03	6.13E-05	1.25E-02	3.54E-04	-1.26917	0.204	0.432
Insomnia	28604731	sleeping	0.063	0.151	0.043	0.052	1.41E-01	3.80E-03	2.72E-01	1.44E-02	1.299458	0.194	0.428
Insomnia	27992416	sleeping	0.077	0.174	0.037	0.048	3.62E-02	3.00E-04	9.10E-02	1.49E-03	1.607508	0.108	0.409
Cigarettes smoked per day	20418890	smoking_behaviour	0.053	0.169	0.068	0.064	4.38E-01	8.00E-03	6.21E-01	2.75E-02	1.248338	0.212	0.432
Former vs Current smoker	20418890	smoking_behaviour	-0.050	-0.260	0.060	0.073	4.11E-01	4.00E-04	6.06E-01	1.95E-03	-2.22544	0.026	0.258
Ever vs never smoked	20418890	smoking_behaviour	0.155	0.108	0.063	0.058	1.36E-02	6.16E-02	4.16E-02	1.38E-01	-0.54251	0.587	0.741
Serumurate overweight	25811787	uric_acid	0.961	0.950	0.122	0.132	2.85E-15	5.49E-13	3.62E-13	4.65E-11	-0.05688	0.955	0.973
Urate	23263486	uric_acid	1.027	1.067	0.150	0.135	6.57E-12	2.69E-15	3.34E-10	3.62E-13	0.195095	0.845	0.933